Emma C. Boettcher. Predicting the Difficulty of Trivia Questions Using Text Features. A Master's Paper for the M.S. in I.S. degree. April, 2016. 70 pages. Advisor: Stephanie W. Haas

In numerous contexts, including community question answering systems, school exams, and trivia competitions, a need to assess the difficulty of questions arises. This study examines what features predict difficulty in the realm of trivia questions, considering features related to readability and the question's topic as potential contributors. Using clues from the game show *Jeopardy!*, the study finds that features relating to a trivia question's length, the inclusion of audiovisual media, and its constituent noun and verb phrases have a significant impact on the clue's difficulty. Based on these findings, this study proposes that finding more nuanced ways to depict the amount of information in a trivia question would lead to further advancements.

Headings:

Text mining (Information retrieval)

Question-answering systems

Text categorization

Readability (Literary style)

PREDICTING THE DIFFICULTY OF TRIVIA QUESTIONS USING TEXT FEATURES

by
Emma C. Boettcher

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2016

Approved by

_____

Stephanie W. Haas

Table of Contents

Introduction

Questions asked by people seeking information take many forms, depending on the asker, the audience, and the information sought. A question that addresses common knowledge may regardless be expressed complexly, if the person asking does not have enough information to communicate precisely. Conversely, a question about a more obscure piece of information may be expressed clearly if the person asking is otherwise an expert. Questions may even be designed to be difficult, as with trivia questions or questions on exams, in order to learn about or reward those who respond correctly (Heilman, 2011).

Though many contexts exist in which questions' difficulty levels might be manipulated, the context explored by this paper is that of writing trivia questions, which require their answerers to recall information but may not require any higher cognitive function. These questions, while they may be addressed toward an individual, known audience, are often addressed to a group audience whose levels of knowledge are mixed or otherwise unknown. Trivia-based board games, quiz nights, and game shows all rely on human estimates of what makes a question difficult or easy for an unknown audience. The game show *Jeopardy!* is used in this study because it publishes and quantifies its

difficulty ratings of trivia clues (what the show terms "answers," though they are more similar to other contexts' trivia "questions"), rewarding contestants with more money if they respond correctly to more difficult questions. In this dataset, the clues are clearly assigned labels as to their difficulty levels, with each label having required an individual human judgment of the clue's difficulty.

Though these human evaluations of difficulty are commonplace, it is more challenging to automatically assess difficulty through machine learning. This problem is most commonly addressed by community question answering systems, though those solve the slightly different problem of matching a user's profile to a question. The system can learn about what questions the user has previously answered, and use that information to recommend recently submitted questions based on the user's perceived knowledge base. However, even in these question answering systems, information about the users is not always available (particularly for new users). In addition, experts prefer to answer questions without good answers because the harder it is to answer a question, the more valuable their answer is (Pal & Konstan, 2010). Giving experts questions that are determined to be difficult relative to the knowledge level of the community, rather than to the knowledge level of specific users, can promote engagement (Pal & Konstan, 2010). Estimating the difficulty of questions with regard to an unknown community can also provide information on the users who answer them by demonstrating the users' behavior when faced with a question that is very difficult or very easy (Liu, Wang, Lin & Hon,

2013). Learning more about what makes a question objectively difficult therefore may have broader implications beyond creating questions for trivia games.

In order to study what features contribute to trivia clue or question difficulty, two possible underlying factors are studied: readability (or the form of a question) and topic (or the content or subject of a question). Readability addresses how much cognitive effort is required to make sense of the form of a question. Literature on using text mining to predict readability suggests that features such as the rarity of a word in a corpus, the syntactic structure of a document (its phrases, clauses, and parts of speech), and the presence of specific vocabulary may contribute toward readability. Difficult syntax, complicated or unfamiliar terminology, and overly long phrases, sentences or documents can make readers perceive a text as difficult. Though media is not often addressed when considering readability, in this study, the use of video, audio, and image files are grouped with readability as it relates to the form (non-verbal) of the trivia question or clue being studied.

Another contributor to the difficulty of a trivia question or clue is the subject of the information need, which is operationalized in this study by assigning clues to topics and determining whether a clue represents a new topic. Novelty detection, new topic detection and other fields provide framework for this endeavor, as newer, more obscure topics may be found in more difficult clues. Examining latent features such as the novelty of the information need can act as a proxy for the obscurity of the topic and therefore the

difficulty level. If a topic is very new, and therefore rarely asked about, the knowledge in it is likely expert level. In contrast, if a topic is asked about frequently, the frequency of the information need perhaps counter-intuitively demonstrates that the need may be at the novice level. In a community question answering system, novices may submit more questions than experts; in the dataset of trivia clues, topics that appear over and over again may reflect the opportunity its audience has had to become familiar with the topic. Recording the topic (or topics) of each trivia clue shows how these topics and their representation in the dataset affect its perceived difficulty level.

Using text mining, this paper addresses whether those two groups of features, readability and topic, can predict whether a trivia clue is considered easy or difficult. Using clues from the game show *Jeopardy!*, which rates the difficulty of its clues on a five-point scale, these features are used to answer the following research questions:

> **Research Question 1**: What text features predict difficulty for fact-finding questions?
> **Research Question 2**: How do features associated with readability contribute to predictions of difficulty for fact-finding questions?
> **Research Question 3**: How does the topic of an information need contribute to predictions of difficulty for fact-finding questions?

Literature Review

Readability

In the field of education, readability can be calculated on an ordinal scale which reflects how children learn to read, such as the Flesch-Kincaid scale. Readability is often therefore only considered in the context of documents for children (Paukkeri, Ollikainen & Honkela, 2013). However, even though a document may not be meant for children, as is the case for clues in the *Jeopardy!* corpus, its text may occupy any number of readability levels. These levels, though, are also interpreted differently throughout the literature. Researchers often propose variations on these readability scales without attempting to precisely pinpoint readability. Liu, Croft, Oh & Hart (2004), instead of testing whether their classifier can correctly predict the exact readability, divided reading levels into two or three broad categories at opposite ends of the scale and evaluated if their classifier could predict the correct category. Further reducing the need for an exact scale, Tanaka-Ishii, Tezuka, & Terada (2010) predicted the difficulty of documents relative to each other without attempting to rank absolute readability. With this in mind, exact readability is not computed for any clue in this study, but the features that predict readability are assessed for each document.

The features used to predict readability in the literature are largely consistent with each other. Liu et al. (2004) divided the features they considered to predict readability into semantic and syntactic categories, exploring both the words and structures of sentences. Semantic features of a document relate to the content and words found in it,

particularly the frequency of *n*-grams in a document in comparison to other documents. Syntactic features, on the other hand, relate to the structure of a document and the words within it. Liu et al. placed sentence length, average number of characters per word, average number of syllables per word, and the distribution of different parts of speech in this syntactic category. Subsequent work attempts to isolate which of these features are the most predictive of readability. Tanaka-Ishii et al. (2010) modeled only a document's vocabulary and ignored length altogether when developing a model that predicted relative readability of two documents. Though their model was successful in predicting relative readability, it did not output absolute readability evaluations. However, Collins-Thompson & Callan (2005) also used unigram features (including stopwords) to predict the absolute readability of a document, obtaining a low root mean squared error.

Feng, Jansche, Huenerfauth & Elhadad (2010) divided their features into six sets: discourse features (such as presence of named entities, synonyms in lexical chains, resolved coreferences), language modeling that removes words with a low information gain, syntactic features (such as number of phrases, average phrase length and average number of phrases per sentence), shallow features (usually relating to words per sentence or number of syllables per word), perplexity features, and part-of-speech features. They divided parts of speech into content words (nouns, verbs, numbers, adjectives, adverbs) and function words to analyze their impact on prediction, observing that part-of-speech features, particularly nouns, are the strongest predictors of readability, though they suggested that discourse level features may be more useful when texts are more complex.

Based on these findings, certain features from the readability literature are included in the study. The impact of unigrams (including stopwords) is studied, though

the threshold for how many documents a unigram needs to appear in to be included as a feature is also evaluated at three levels, in order to capture somewhat how the rareness of a word affects difficulty predictions. (As seen in the next section, the unigrams are assigned binary weights, though inverse document frequency weights would likely also prove fruitful.) In addition, features relating to the length of the clues are used, as in Liu et al. (2004), and features relating to noun and verb phrases, since nouns are considered the most salient parts of speech for predicting readability.

New event detection

In addition to considering readability features, this study also examines how the presence of new topics or new information affects difficulty predictions. Measuring a new fact in a corpus is analogous to detecting a new event in a news stream. In both instances, a document is being compared to what already exists in order to determine the amount of new information it contains. Most new event detection systems use temporal features and sudden increases of relevant information (text "burstiness") as features to inform the system, but these features may be less useful when analyzing fact-finding questions, particularly fact-finding questions which are designed to be difficult and not repeat answers too often, as in the *Jeopardy!* corpus (McCown, 2015). However, new event detection systems also measure similarity between topics. In addressing news streams, Kumaran & Allen (2005) used cosine similarity based on names, topics, and the entire document to predict whether a document represents a new event. They concluded that this model performs well for predicting whether a document is part of an older news story, though it is less successful for predicting a new topic. Topics are therefore considered somewhat useful for predicting the newness of a document.

Novelty detection and similarity

Novelty detection is a specific problem that looks at whether a document contributes new information to existing documents. Usually expressed in terms of a binary classification (a document either contributes new information or it does not), the task has seen progress despite the fact that human annotators struggle to agree on what constitutes new information. Otterbacher & Radev (2006) found that human annotators could usually agree on what constituted relevant information for a topic but agreed less often about whether a piece of information constituted new information for a particular information need. This apparent disparity between topic and information is reflected in the balance of research; more has been done about topic similarity and new topics than new information itself, suggesting that modeling a clue's topic would be more fruitful for this study than addressing whether it contains new information.

As in detecting readability, unigrams, bigrams, and syntactic features have been used to determine the similarity of documents (Karampatsis, 2015; Diao, Xu & Xiao, 2011; Ng, Tsai, Chen & Goh, 2007; Zhang & Tsai, 2009; Eyecioglu & Keller, 2015). Tsai & Kwee (2011) compare the different methods of weighting of *n*-grams, and conclude that while TF.IDF weights are the best predictors for sentence-level novelty, binary weights perform better when judging the novelty of documents if there are not many novel documents in a set. Named entities are also somewhat effective predictors of topic similarity, but authors using these features acknowledged that the approach did not match the state of the art (Diao et al. 2011, Karamapatsis 2015, Ng et al., 2007; Zhang & Tsai, 2009).

More success has been found in augmenting the vocabulary used in a document in order to increase the accuracy of predictions of semantic similarity. Two major toolkits exist: the Stanford GloVe system, which computes the cosine similarity between word vectors and automatically determines their similarity, and WordNet, a manually created corpus that defines synonyms and hypernyms (that is, words of a which a given word is a subset) of each word the corpus (Pennington, Socher & Manning, 2014). GloVe system is considered state-of-the-art for detecting paraphrases. Karampatsis (2015) used GloVe to evaluate existing potential paraphrases was able to accurately predict when one Tweet paraphrased another. Similarly, Sanborn & Skryzalin (2015) used the Stanford NLP GloVe system to evaluate the semantic similarity of statements. However, in research more closely related to this study, WordNet has been used to create paraphrases, instead of simply evaluate them. Petrović, Osborne & Lavrenko (2012) used WordNet to paraphrase their documents, considering all words in a synset as candidates to paraphrase any given word. This method improved over a baseline method to detect new events from documents. Li & Manandhar (2011) combined WordNet paraphrases with unigram features in order to measure similarity in question answering systems. Paraphrasing documents, no matter what toolkit, allows researchers to look at the information need itself instead of the language used to express it. For this paper, creating paraphrases allows the distinction between readability of the question (operationalized through unigrams) and newness of the underlying information need.

Two methods for computing document novelty dominate the literature. Cosine-similarity examines a document and compares it to previous documents that have been determined to be similar in the same way. Documents are represented as vectors, with

individual words occupying dimensions and being weighted by term frequency, term frequency-inverse document frequency (TF.IDF), or even binary weights. This approach is computationally expensive to do for all documents in a large corpus (Karkali, Rousseau, Ntoulas & Vazirgiannis, 2013). The alternative is to model the topics of the documents using a probabilistic model and then compare the topics of new documents to those topics (Aksoy, Can & Kocberber, 2012; Wanas, Magdy, & Ashour, 2009).

Question answering

Community question answering is the area of research most directly connected to determining the similarity and novelty of fact-finding questions. Much of the literature focuses on the community and social aspects of these question answering systems, exploring the link structure between questions, the affirmation received by users from other community members, and the multiple answers given to a question (Yang, Qiu, Gottipati, Zhu, Jiang, Sung & Chen, 2013). However, as informative as these features are, textual features have been shown to be valuable predictors of various behaviors on these sites. Burel, Mulholland, He & Alani (2015) considered both user features and question features when studying community question answering communities. In predicting answering behavior, Burel et al. addressed features such as polarity and age of question that are less relevant in the corpus to be studied, but finds that answering behavior is mostly determined by question features as opposed to user features, with readability being the fourth most predictive feature. Similarly, Liu et al. (2013) addressed how community participants had scored answers to evaluate the difficulty of questions, but noted a correlation between this and text features. These results show that while the focus

in community question answering research is the representation of users, the representation of text may be as important.

Many studies on community question answering systems use topic modeling to detect a new question's similarity to older ones. Duan, Cao, Lin & Yu (2008) proposed modeling the similarity of questions based on their topic and focus. With a similar motivation, many studies use the categories that have been assigned to the question as a feature or as a criterion for feature selection (Cai, Zhou, Liu, & Zhao, 2011; Zhou, Chen, Zeng & Zhao, 2013; Zhou, Li, King, Lyu, Song, & Cao, 2011). Nomoto (2010), while working in novelty detection, addressed the information a document's collection provides in addition to the information in a document itself, showing that information from both the collection and the document can assist in detecting similarity. Adding additional information to reduce the number of features and address the focus of the question specifically is important for long queries, but short queries need to be augmented with other information (Zhou, Lai, Liu & Zhao, 2012).

This additional information can come from many sources. Drawing directly from the corpus, existing questions' features can be expanded by looking the answers they have received. While acknowledging that this information is not available for new questions, both Ji, Xu, Wang & He (2012) and Dror, Koren, Maarek & Szpektor (2011) incorporated answers while modeling the topics of questions. Knowing that named entities contribute to topic modeling, Singh (2012) used an entity catalogue in order to build on topic models. Also as in topic modeling, paraphrases can be used in order to ascertain semantic similarity. Bunescu & Huang (2010) described three levels of question relatedness in the context of searching for redundant questions: paraphrase (semantically

identical), useful (relevant), and neutral (not related to the information need). In this case, both paraphrases and useful questions provide information on the novelty of the question in the corpus.

With regard to these finds, this study assigns its clues to topics created using other clues not in the dataset, and creates and assigns topics based on a clue's text, its category, and the correct response. In addition to using this text, the study also tests three different variants on building topic models: topic classification using the original clue text, topic classification using the original clue text expanded with hypernyms, and topic classification using the original clue text expanded with synonyms. These hypernyms and synonyms are obtained using WordNet, since it has been previously used to paraphrase documents to detect topic similarity. A clue's membership in topics and the number of topics it could plausibly belong to are used as approximations for the clue's overall obscurity.

Though not discussed to the same extent as topic in the community question answering literature, multimedia questions have also been shown to affect behavior in those communities. Carroll (2015) found a correlation between the presence of media in questions and the number of useful answers a question received; however, it is unclear whether this correlation emerges from multimedia questions being easier to answer or if they simply appear more attractive to the answerers of a community. Regardless, the presence of media in trivia questions is explored in this study to see if Carroll's research can be applied to this arena. However, media is grouped with the readability features discussed above, since it more closely relates to the form of a clue than the topic of it.

Trivia

Question answering systems may comprise a variety of questions, but this study only addresses trivia questions, using specifically clues from the game show *Jeopardy!* While the nuances of this dataset are discussed in the Methods section, it is worthwhile to discuss how scholars and trivia writers alike differentiate levels of difficulty. A common taxonomy of cognitive processes separates learning objectives into six levels: remembering, understanding, applying, analyzing, evaluating, and creating (Krathwohl, 2002). Remembering requires a person to recall details or facts, but understanding involves making inferences from established facts, among other processes. Both of these processes appear in the dataset, and are discussed below, but the higher levels of learning objectives discussed by Krathwohl are more complicated than the ones summoned when responding to trivia questions.

These two simplest processes, however, are further broken down by trivia experts and writers. Jennings (2007) described a nine-part taxonomy of trivia, of which the key attributes are how answerable and entertaining a question or clue may be. Though some of these categories overlap, Jennings (2007) isolated three categories of interest for this study: the plain vanilla recall, plain vanilla with hot fudge, and the puzzler. The baseline for a trivia question or *Jeopardy!* clue, Jennings explains, is the "plain vanilla recall," where the answerer will either know the answer or not (Jennings, 2007, p. 113). One example is the following clue, which appeared on *Jeopardy!* in 1989:

AIRPORT DESIGNATIONS $1000: DTW (Detroit)[1]

[1] This paper formats clues as they appear in online *Jeopardy!* archives and discussion forums, using the template <CATEGORY> <value>: <clue text> (<correct response>). Where necessary, I have clarified the difficulty level of the clue being discussed within the text of the paper.

The simplicity of these questions can easily become repetitive or boring, so where trivia is meant to entertain, the clue can be expanded with inessential facts - the "hot fudge" to the original question's vanilla (Jennings, 2007, p. 113). Seventeen years later, the same information as above was used again on *Jeopardy!*:

> AIRPORT CODES $1000: Its McNamara Terminal is also the Northwest Airlines World Gateway: DTW (Detroit)

A few might know the McNamara Terminal or the Northwest Airlines World Gateway, but this additional information mainly prevents repetition and makes the clue less dry. Jennings did not discuss the impact of this additional information on the clue's level of difficulty, though according to the show, the clue is considered to be the most difficult of its category (as indicated by its being worth $1000) regardless of the additional information provided.

Most of Jennings' taxonomy further focuses on these factual recall clues that fit into the first level of Krathwohl's taxonomy, but the last category Jennings described is that of the "puzzler," which "*nobody* knows the answer to" but can be deduced with enough peripheral trivia knowledge and a logical inference (Jennings, 2007, p. 116). Most closely corresponding to the "understand" level of Krathwohl's taxonomy, this plays out in *Jeopardy!* clues such as the following:

> $800 GO GREYHOUND: Dedicated to the adoption of ex-racing greys, First State Greyhound Rescue places dogs in N.J., Penn. & this state (Delaware)

First State Greyhound Rescue, according to this clue, only operates in three states, so the likelihood that any contestant would have heard of it and know the extent of its operations is small. Regardless, this clue is considered "easy" because of the phrase "First State," which evokes the first state in the United States, Delaware. However, a

person still has to know what the first state is in order to respond to the clue correctly, so the clue is predicated both on the "remember" and "understand" levels of cognition. While trivia mainly focuses on remembering pieces of information (with or without bonus material accompanying the main part of the clue) it does also occasionally require people to make inferences from the information they can recall, especially where the goal is to entertain.

Another element of constructing clues that Jennings and other trivia writers acknowledge is that the same event or fact can be referenced through different levels of obscurity, and therefore two clues with the same response can have different levels of difficulty. Will Shortz, puzzle editor for *The New York Times*, can take a completed crossword grid and reverse-engineer the clues for three different levels of play, so that beginners, intermediates and experts will use different sets of clues to fill in the same letters in the puzzle (Romano, 2005). Clues therefore are acknowledged as the real indicators of difficulty, not the content of the responses they require (Gaffney, 2006; Romano, 2005). Crossword construction has its own constraints, but there are underlying similarities to *Jeopardy!*: both have an imagined, broad audience and both need to assess the difficulty of their material. In both, the way the clue is phrased has more to do with its difficulty than the answer, which makes the text of the clue, not the correct response, the main subject for this study.

Methods

The research questions for this paper are restated here:

> **Research Question 1**: What text features predict difficulty for fact-finding questions?
> **Research Question 2**: How do features associated with readability or the form of a question contribute to predictions of difficulty for fact-finding questions?
> **Research Question 3**: How does the topic of an information need contribute to predictions of difficulty for fact-finding questions?

In order to answer the above research questions, I use clues and responses from the game show *Jeopardy!* Clues within the game show are assigned into categories and within those categories are assigned a difficulty level. Given a clue (what the show calls an "answer"), the contestants must come up with a correct response (in the show's terminology, a "question"). Clues are divided among three rounds: the Jeopardy! round, the Double Jeopardy! round, and the Final Jeopardy! round. The Jeopardy! and Double Jeopardy! rounds comprise thirty clues each, divided into six categories. Each category has five clues, and the five clues are assigned a dollar value according to the clue's difficulty (McCown, 2015). The easiest clues are valued at $200 in the Jeopardy! round and $400 in the Double Jeopardy! round; the most difficult clues are valued at $1000 in the Jeopardy! round and $2000 in the Double Jeopardy! round. Though most clues are assigned a value, several clues throughout the game allow contestants to choose how much money they want to wager on a clue, as in the Daily Doubles, which occur in the first two rounds, and the clue in the Final Jeopardy! round. Unlike Daily Doubles, which

are assigned a value before contestants wager on them, the Final Jeopardy! clue is not assigned a value and therefore has no implied difficulty level.

Clues from *Jeopardy!* are available from the J! Archive.[2] As of March 13, 2016, the J! Archive consists of 301,463 clues, which are organized into 5,211 games as they originally appeared on the show. The show number and airdate are recorded for each game. A clue's text, correct response category, original value, and actual value (if it is a Daily Double) is recorded, as well as information about which contestants responded to the clue and if they did so correctly. Within the clue's text, there may also be links to media materials that were presented with the clue and are stored in the J! Archive.

In January 2014, the J! Archive was scraped, and a file of 216,930 clues from 3,640 games was posted on Reddit. In this scrape, the text of the clue (including links to media materials) is recorded, along with the show number, airdate, category, and correct response, as shown in Table 1.

Date	Round	Category	Value	Clue	Response
3/16/11	Jeopardy!	ACTING OUT	$400	It's the pointless thing Sarah is doing	comparing apples to oranges

Table 1. Data available from Reddit scrape for an example clue.

Though the scrape does not include all data from the J! Archive, it contains sufficient information and its convenient format made it an ideal source of clues for the study.

From this scrape, a narrower selection of clues was used as the dataset for this study. The selection needed to capture a uniform selection of clues, while still maintaining a large enough scope so that patterns of difficult clues or obscure topics could emerge. In order to achieve these goals, the dataset consisted of clues from only

[2] http://www.j-archive.com/

two seasons (from 9/14/2009 to 7/29/2011) from the scrape. Including multiple seasons worth of clues allowed for the possibility that topics could be reused, but limiting the dataset to two seasons minimized any variations in clue writing or language that might have occurred over the course of the show.

After the two seasons were selected from the scrape, additional clues were removed from the dataset to make it more uniform. Within those two seasons, a number of tournaments were held for specialized contestant groups (including teens, college students, and returning multi-day champions), but clues from these groups were also removed from the dataset in order to minimize variations in clue difficulty assignments. Similarly, clues that did not require the contestant to remember information but instead solely required the contestant to perform a higher cognitive function were also removed to make the assignments of clue difficulty more consistent. (A list of these clues, all of which involved solving math problems, can be found in Appendix A.) Finally, clues that originally had had no difficulty value assigned to them (those in the Final Jeopardy! round and end-of-game tiebreaker clues) were also removed from the dataset.

Of the remaining clues, several steps were taken to normalize clue values. Daily Doubles, which were represented in the original scrape of the archive by the amount a contestant wagered on them, were reassigned their original clue value. Also, though the clue values differ between the Jeopardy! round and the Double Jeopardy! round, clues of corresponding values between rounds were assumed to be of the same level difficulty. As shown in Tables 2 and 3, the clues were assigned into five nominal categories: very easy, easy, medium, difficult, very difficult.

Difficulty level	Jeopardy! round	Double Jeopardy! round	Number of clues in dataset (%)
Very easy	$200	$400	4,416 (20.2%)
Easy	$400	$800	4,393 (20.1%)
Medium	$600	$1200	4,384 (20.1%)
Difficult	$800	$1600	4,350 (19.9%)
Very difficult	$1000	$2000	4,292 (19.7%)

Table 2. Difficulty levels of clues and their proportion in the dataset.

Date	Round	Category	Value	Clue	Response	Difficulty
3/16/11	Jeopardy!	ACTING OUT	$400	It's the pointless thing Sarah is doing	comparing apples to oranges	easy

Table 3. Example clue with its assigned difficulty level.

Before extracting features, non-clue text was removed from the clues. Any links were removed, though the type of materials they linked to was preserved as metadata for the clue. Links to media files were occasionally accompanied by descriptions of their content; these descriptions were also removed because those descriptions were added by archivists and not originally presented to contestants as part of the clue. Lastly, while most of the file formats were taken to be accurate representations of the clue as it originally appeared, the archive often uses only images (.jpg files) to represent video clues (.wmv files) when the video does not provide additional information. Parenthetical statements at the beginning of clues describing the person giving the clue usually indicate that the clue was a video, so if these parenthetical statements link to a .jpg file, it is

recorded in the metadata for this study as a .wmv file. An example clue and its media values are shown in Table 4.

Clue	Response	Difficulty	wmv	jpg	mp3
It's the pointless thing Sarah is doing	comparing apples to oranges	easy	1	0	0

Table 4. Media values and cleaned text for an example clue.

Feature generation

Once the clues were cleaned and their values were normalized, features for each clue were generated. Five groups of features were generated for each clue: unigrams, length, media, phrase, and topic. As described in the literature review, these features were thought to have some correlation with the difficulty of a clue.

Unigrams were generated using the LightSide text mining software.[3] Three versions of the set of unigrams were created based on the number of clues a unigram appeared in within the dataset:

- $unigrams_1$: All unigrams that appeared in at least one clue (in effect, all unigrams in the dataset)
- $unigrams_5$: All unigrams that appeared in at least five clues, but excluding unigrams that only appeared in four or fewer clues
- $unigrams_{10}$: All unigrams that appeared in at least ten clues, but excluding unigrams that only appeared in nine or fewer clues

For each version of unigrams, the unigrams were assigned binary weights. Since the clues are short, the frequency of a given term in any clue is relatively small, so term frequency weights were not considered. Though inverse document frequency weights would have

[3] http://ankara.lti.cs.cmu.edu/side/

indicated a word's rareness in the corpus, binary weights were judged to be sufficient based on the literature.

Since each clue may contain references to media files, the references were stored as metadata for the clue, as previously described. Occasionally these references included descriptions of what was present in the media (e.g., "Kelly of the Clue Crew shows a map on the monitor" to describe the content of a video), but these descriptions were removed so that the text of the clue only consisted of the actual wording given to the contestants. The number of pictures (.jpg files), videos (.wmv files), and audio files (.mp3 files) presented to the contestants within clue was stored for each clue.

Length was also considered as a predictor of clue difficulty. The length of each clue was measured in three ways: characters, syllables, and words. Syllables were counted using the Carnegie Mellon University Pronouncing Dictionary in the Python Natural Language Toolkit (NLTK). This dictionary assigns the second halves of contractions their own syllables, though as in the following example, this adds an extra syllable to the sentence.

It • 's • the • point • less • thing • Sar • ah • is • do • ing → 11 syllables

Words not in the CMU dictionary were handled by counting the number of non-consecutive vowels in the words as an approximate measure of the number of syllables. While this meant that initializations such as FDR or JFK were counted as having 0 syllables, the syllable counter had an overall accuracy of 63.6% when tested on 500 clues not in the dataset. In addition, numbers were not included in the syllable count, because there are often multiple ways to voice them (the number "2100" can be expressed as "twenty-one hundred" or "two thousand, one hundred"). Numbers also do not have

synonyms; where the use of a polysyllabic word reflects the choice not to use a monosyllabic word (and vice versa), the same cannot be said for numbers. However, to compensate for the numbers not being in the syllable count, the number of numbers in each clue was stored as a separate feature, along with the number of words and the total count of words and numbers. Words were counted using spaces to determine the beginnings of new words; contractions such as “can’t” and “it’s” were counted as two words.

It • 's • the • pointless • thing • Sarah • is • doing → 8 words

The total number of syllables and the total number of characters were also both divided by the number of words and the number of words and numbers, respectively, to calculate the average syllables and characters per word. Length features for an example clue are shown in Table 5.

Clue	**Char.**	**Syl.**	**Words**	**Numbers**	**Words and numbers**	**Char. per word**	**Syl. per word**
It's the pointless thing Sarah is doing	39	11	8	0	0	4.75	1.375

Table 5. Length features for an example clue.

The last readability features extracted were those that related to the noun and verb phrases of the clue. Each clue was chunked into noun and verb phrases using a part-of-speech tagger and regular expression parser available in Python NLTK. After tagging 250 clues not in the tested dataset with their parts of speech, I generated a series of regular expressions that corresponded to the existing noun and verb phrases. (These regular expressions can be found in Appendix B.) The clues in the tested dataset were then

tagged with their parts of speech, and these representations were used to chunk the noun and verb phrases of each clue using the regular expressions, as shown in Figure 1.

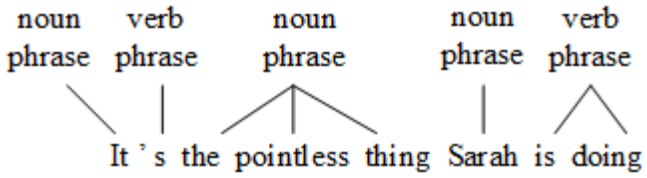

Figure 1. Noun phrases and verb phrases for an example clue.

Once each clue had been chunked into its noun and verb phrases, numeric data about these phrases was stored as features for the clue, including the number of noun and verb phrases, the average length in words of these phrases, and the total number of phrases (noun, verb, and other). Table 6 shows the phrase features for an example clue.

Clue	Noun phrases	Avg. length of noun phrases	Verb phrases	Avg. length of verb phrases	Total phrases
It's the pointless thing Sarah is doing	3	1.666666667	2	1.5	5

Table 6. Phrase features for an example clue.

In addition to readability features being generated for each clue, the topic of a clue was also considered as a potential indicator of difficulty. Since each clue is already assigned to a category and since *Jeopardy!* contestants must rely on the category to estimate what unseen clues are about, the categories were treated as reliable predictions of topic. However, category names often appear only in a single game, and each category has a clue assigned to each difficulty level within it. Using the category names as topics would therefore result in 4,244 topics, all of which would have roughly the same amount of clues from each difficulty level and would be unable to provide helpful information for

a classifier. Rather than treat each category name as its own topic, a method for discovering larger topics that could apply to clues in multiple categories was devised.

In order to discover larger topics existing in the dataset, the category names of the *Jeopardy!* corpus were studied for underlying patterns. Using a subset of 4,084 clues not in the tested dataset, all unigrams, bigrams, and trigrams appearing in category titles over the span of at least 2 games were generated. The least ambiguous of these phrases (for example, "movie_characters" and "book_characters" are less ambiguous than simply "characters") were then grouped into 65 topics using WordNet,[4] Library of Congress subject headings,[5] and relationships in category names to measure similarity. (These unigrams, bigrams and trigrams can be found in Appendix C.) The categories of 22,185 clues not in the tested dataset were then searched for the unigrams, bigrams and trigrams that had been grouped into the topics, and clues whose categories matched those phrases were assigned to their corresponding topics. In this way, each topic had constituent clues that described how the topic manifested itself in the *Jeopardy!* corpus.

Next, these clues were used to create a model for each topic. Clues assigned to a particular topic were combined with an equal number of clues that had been assigned to other topics in order to build a training and test set on which to build the models. Twenty percent of this data was set aside for testing; the rest was used to build a model to predict the topic. After a model was built on the training data using LightSide, it was tested on the remaining data. (A list of the generated models and statistics about their accuracy can be found in Appendix D.) This was repeated for the 65 different topics.

---

4 https://wordnet.princeton.edu/wordnet/

5 https://www.loc.gov/catdir/cpso/lcco/

Each topic model was then used to predict the topics of clues in the tested dataset, using the category names and responses as well as the clue text. Three versions of the clue text were used to assign the clues to topics. One version was the original clue text, one expanded the clue text using hypernyms from WordNet, and one expanded the original clue text using synonyms. An example is shown in Table 7.

Clue version	**Clue text**
Original clue text	It's the pointless thing Sarah is doing
Hypernyms	It's the pointless thing Sarah is doing situation state_of_affairs
Synonyms	It's the pointless thing Sarah is doing unpointed make do

Table 7. Clue text expanded with hypernyms and synonyms for an example clue.

(No attempt was made to incorporate the hypernyms and synonyms naturally into the clue, since the model did not incorporate the position of words into its predictions.) The model for each topic independently predicted topic membership using each version of the clue text, as well as the clue's category and correct answer. Membership in any topic was represented as a binary feature, though each clue could belong to multiple topics. Three versions of topic features were created:

- $topics_{ORIG}$: the predictions made using the clue text as it was originally stated, as well as a clue's category name and response
- $topics_{HYP}$: predictions made using clue text expanded with the hypernyms of any nouns or verbs, as well as a clue's category name and response
- $topics_{SYN}$: predictions made using clue text expanded with the synonyms of any nouns or verbs, as well as a clue's category name and response.

The number of topics a clue belonged to was also a feature, with integer values ranging from 0 to 65. A summary of these features, along with the readability features, is shown in Table 8.

Feature group	Feature set	Component features	Versions
Readability	Unigrams	unigrams	$unigrams_1$ $unigrams_5$ $unigrams_{10}$
	Media	number of pictures number of video files number of audio files	N/A
	Length	number of characters number of syllables number of words number of numbers number of words and numbers characters per word syllables per word	N/A
	Phrase	number of noun phrases number of verb phrases average length in words of noun phrases average length words of verb phrases total number of phrases (noun, verb, and other)	N/A
Topic	Topic	membership for 65 topics total number of topics clue belongs to	$topics_{ORIG}$ $topics_{HYP}$ $topics_{SYN}$

Table 8. Features used in study. Another version of this table, which details the potential values for each feature, is shown in Appendix E.

Experiments

To determine which sets of features were predictors of difficulty, several experiments were run using Weka.[6] Each experiment compared the accuracy of naïve Bayes classifiers built using different feature sets, determined through ten-fold cross validation. In ten-fold cross-validation, the data is partitioned into ten folds, then a model is created by training on nine folds of the data and testing on the tenth, held-out fold. This is repeated ten times, with a different fold being held out each time. In these experiments,

[6] http://www.cs.waikato.ac.nz/ml/weka/index.html

the data was partitioned into ten folds by the date of each clue, in order to prevent clues in the same category from appearing in different folds.

After each model was built using nine folds, the accuracy for each model was recorded and then compared to other models built on the same combination of folds with different feature sets using Fisher's randomization test. Fisher's randomization test evaluates whether the accuracy associated with one fold for a particular set of features represents a significant difference from the accuracy associated with that fold for a different set of features (Smucker, Allan & Cutterette, 2007). To calculate the statistic, the mean accuracy of the ten folds is calculated for both sets of features being evaluated, and the difference between those two means is recorded. Then, for each fold, the accuracy from the two models is randomly assigned or reassigned to either set of features. The mean accuracy for each set of features' folds (now with some accuracies switched between the two groups) is computed again and the difference between them is recorded. This is repeated 100,000 times in order to create an adequate sample of the existing random permutations of the results. The number of times the difference between the mean accuracies is greater than or equal to the original mean is divided by 100,000 to obtain the probability that the original difference between the two models was due to chance. If the p-value was less than .05, the difference between the models was considered significant.

Using this procedure, a number of different comparisons between groups of features were evaluated. First, the optimal set of unigrams was determined by comparing the three versions of unigram features: $unigrams_1$, $unigrams_5$, and $unigrams_{10}$. In order to determine the best version, the accuracy of a model built using $unigrams_{10}$ was compared

to the accuracy of a model built using $unigrams_5$. Then, if a significant difference was found, the model with the highest accuracy was compared to the models built on $unigrams_1$. If a significant improvement in accuracy was made according to Fisher's randomization test, the more accurate model was used for the next stage of the study. If in either comparison, there was no significant difference in accuracy between the two models, the model with a smaller feature set was used for further experiments because it was more efficient. A summary of these experiments is shown in Figure 2.

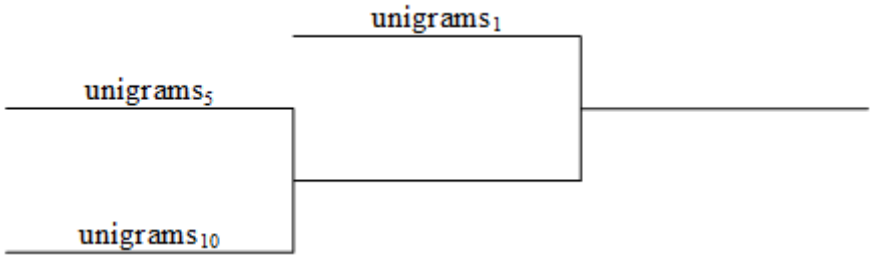

Figure 2. Plan for comparisons between different versions of unigrams.

Evaluating what kinds of features had significant impacts on accuracy was broken down into two feature ablation studies. The accuracy of the folds of the model built using all of the readability features (length, media, phrase, and the best-performing version of unigrams) was compared to the accuracy of the folds of models built without one set of features. The difference between the models' accuracies was evaluated for significance using Fisher's randomization test. A significant decrease in accuracy in a model that was not built using a particular set of features would indicate that the set of features was useful to the model in predicting difficulty.

The sets of readability features that had a significant impact on accuracy were used to create a baseline for the topic experiments. For each group of topic features (generated from only the clue text, from the clue text and hypernyms, and from the clue text and synonyms), a naive Bayes classification model was built with that group of topic

features and the successful combination of readability features. As above, a ten-fold cross-validation process was used, and the accuracy of each fold was compared the accuracy of the corresponding fold in the model built using only the significant readability features. A significant decrease in accuracy compared to the model that incorporated topic features, as judged by Fisher's randomization test, would indicate that the topic features were useful to the model in predicting difficulty.

In addition to the experiments run above, models were built using a naive Bayes classifier and individual groups of features, in order to study how the model was using each group of features and the confusion matrix associated with each. These models were also built using ten-fold cross validation, but their accuracy was not evaluated for significance.

Results

Unigram threshold

Three models were created, each using a different group of unigrams: all the unigrams in the dataset ($unigrams_1$), all the unigrams appearing in at least five clues in the dataset ($unigrams_5$), and all the unigrams appearing in at least ten clues in the dataset ($unigrams_{10}$). The accuracy and results from this experiment are shown in Table 9.

Fold	**Accuracy of $unigrams_1$ (%)**	**Accuracy of $unigrams_5$ (%)**	**Accuracy of $unigrams_{10}$ (%)**
1	20.5448718	20.7188645	20.5265568
2	20.1144689	19.9313187	20.6082418
3	20.242674	20.1053114	20.7692308
4	20.3846154	20.5631868	20.8608059
5	20.6410256	20.746337	20.8928571
6	20.0778745	20.1282639	20.123683
7	20.4764086	20.7100321	21.2001832
8	19.8579936	19.9038021	20.1191022
9	20.4580852	20.4489235	20.8474576
10	19.7938617	19.8534127	19.9954191
Mean accuracy	20.2591879	20.3109453	20.5943538
p (when compared to $unigrams_{10}$)	**.00198**	**.00837**	N/A

Table 9. Accuracy and p-values for unigrams experiments.

Using Fisher's randomization test, the probability that improvements in accuracy between $unigrams_{10}$ and $unigrams_5$ was determined to be significant (p = **.00837**). The accuracy of $unigrams_{10}$ and $unigrams_1$ was then compared and also determined to be statistically significant (p = **.00198**). Because the $unigrams_{10}$ was determined to be a significant improvement over the other versions of unigrams, $unigrams_{10}$ was used for later experiments.

Readability features

Five models were created: one with all sets of readability features, and four others which represented the total group of readability features without one of the member sets - $unigrams_{10}$, media, length, and phrase. The accuracy and results from this experiment are shown in Table 10.

Fold	**Accuracy of $unigrams_{10}$, media, length, phrase (%)**	**Accuracy of media, length, phrase (%)**	**Accuracy of $unigrams_{10}$, media, phrase (%)**	**Accuracy of $unigrams_{10}$, length, phrase (%)**	**Accuracy of $unigrams_{10}$, media, phrase (%)**
1	21.2454212	21.9276557	21.2728938	21.1401099	20.8882784
2	21.1813187	21.8406593	21.0531136	21.0989011	20.6822344
3	21.3507326	21.4239927	21.2820513	21.2637363	20.728022
4	21.6620879	21.4331502	21.5979853	21.4331502	21.0943223
5	21.5659341	21.3553114	21.5201465	21.2912088	21.0668498
6	21.2414109	21.6582684	20.9986257	21.1726981	20.6550618
7	21.7819514	21.8827302	21.7773706	21.7361429	21.3742556
8	20.8612002	21.7544663	20.6138342	20.6367384	20.3206596
9	21.3559322	21.7590472	21.3055428	21.3559322	21.0627577
10	21.0032066	21.6994961	20.80623	20.8886853	20.2885937
Mean accuracy	21.3249196	21.6734777	21.2227794	21.2017303	20.8161035
p (compared to $unigrams_{10}$, media, length, phrase)	N/A	.98682	**.00311**	**.00177**	**.00104**

Table 10. Accuracy and p-values for readability experiments.

Using Fisher's randomization test, the probability that the differences between the models not incorporating media, length, or phrase features and the model incorporating all sets of readability features was determined to be less than .05, and therefore significant.

Since the media, length and phrase features were all determined to be significant, they were used to create a baseline model, readability, for the topic experiments. Because $unigrams_{10}$ did not have a significant impact on accuracy (in fact, the model not

incorporating $unigrams_{10}$ had a better mean accuracy than the model that did), this set of features was not used in the topic experiments.

Topic features

Four different models were created to evaluate the impact of topic on difficulty predictions. The first combined only the significant readability features (media, length, and phrase features). The other three models that combined the significant readability features with different versions of the topic set of features. One combined the significant readability features with topic membership based on the original text of the clue ($topics_{ORIG}$), another combined the significant readability features with topic membership based on the text of the clue expanded with hypernyms ($topics_{HYP}$), and the last combined the significant readability features with topic membership based on the text of the clue expanded with synonyms ($topics_{SYN}$). The accuracy and results from this experiment are shown in Table 11.

Fold	**Accuracy of readability (%)**	**Accuracy of readability, $topics_{ORIG}$ (%)**	**Accuracy of readability, $topics_{HYP}$ (%)**	**Accuracy of readability, $topics_{SYN}$ (%)**
1	21.9276557	21.5705128	21.753663	22.0100733
2	21.8406593	21.6575092	21.5567766	21.6758242
3	21.4239927	21.7078755	21.6529304	21.9505495
4	21.4331502	21.6071429	21.7857143	21.4377289
5	21.3553114	22.1565934	21.959707	21.996337
6	21.6582684	21.7315621	21.7682089	22.1438388
7	21.8827302	21.3284471	21.424645	21.2551535
8	21.7544663	21.6765918	21.7269812	21.639945
9	21.7590472	22.2766835	21.7224004	21.5849748
10	21.6994961	22.0568026	21.8644068	21.7132387
Mean accuracy	21.6734777	21.7769721	21.7215433	21.7407664
p (compared to readability)	N/A	.22847	.32147	.30026

Table 11. Accuracy and p-values for topic experiments.

Each model built using the topic features was compared to the model that only used significant readability features using Fisher's randomization test. However, none of them achieved a p-value less than .05, so no version of topic features had a significant impact on accuracy.

Discussion

Overall, many of the features hypothesized to have a significant bearing on predicting difficulty were in fact useful in predicting clue difficulty. However, though media, length and phrase features had some impact on accuracy, it was still very slight, and unigrams and topic did not affect accuracy significantly, contrary to what was suggested by the literature. Below, I discuss the potential reasons behind the significance of results.

The significant improvement in accuracy after decreasing the number of unigrams used is not surprising. Including unigrams that have appeared at least ten times in the dataset in the training set prevents the model from assigning importance to relationships between unigrams and difficulty classes that may only occur by chance. For example, the unigram “Wozniak” only appears twice in the dataset; once in the very easy class, and one in the difficult class. If these instances are split between the training and test sets during cross-validation, the model will use the information from the training set to wrongfully predict the difficulty of the clue with that unigram in the test set. However, if the set of unigrams is limited to only those that appear in at least ten clues in the dataset, those ten instances provide a fuller picture of which classes the unigram is frequently associated with.

Though this conclusion is in line with many findings in text mining literature, it is also slightly surprising in this context because the rareness of words can signal a more specialized, esoteric level of vocabulary, which can indicate that a clue is more difficult

to read and therefore in a higher level of difficulty than other clues. Future research in this regard might involve using inverse document frequency weights for unigrams, but those weights might also generate wrong predictions on the test set. Though the 21,834 clues used had only 33,494 types (including words, numbers, and punctuation), some of the hapax legomana comes from misspellings (either from typos made in data entry or categories where contestants are required to correctly spell misspelled words) and anagrams, where the contestant is not required to make meaning from the words but instead assess their components. The following two clues show how a word may not just have multiple meanings, but multiple cognitive interpretations:

> ANAGRAMMED NOTABLES $2000: An ancient playwright: Pi residue (Euripides)
> ACCIDENTAL INVENTIONS $1200: In 1879 a chemist's unwashed hands got residue on a roll, which tasted sweet; he'd patent this, 300x sweeter than sugar (saccharin)

In the latter, “residue” is a word that carries meaning; in the former, “residue” is only an arrangement of letters.

This ambiguity between unigrams that convey meaning and unigrams that only combine letters may have somewhat contributed to the fact that unigrams are not significant in predicting difficulty, but there were likely other factors. Among these is the idea that each clue might contain a mix of information; instead of being composed of only “easy” unigrams or only “difficult” unigrams, there are mixes of unigrams in each clue which present mixes of “easy” or “difficult” trivia. For example, a clue that the unigrams-only classifier correctly identified as “easy,” two entities are mentioned:

> THE TENTH INNING $400: You can't talk about baseball without mentioning this small Caribbean nation that's given us players like Jose Reyes & Sammy Sosa (the Dominican Republic)

In *Jeopardy!* clues, Jose Reyes only appears one other time, in a clue that is classified as "difficult." However, Sosa appears in other clues marked "easy" or "very easy." The identification of this clue as "easy" by the writers may therefore be more dependent on Sammy Sosa than it is on Jose Reyes, making Sosa a more meaningful unigram than Reyes. This idea suggests that the unigrams of a clue do not combine equally to create a clue's difficulty, but that easier unigrams can outweigh more difficult ones. In addition, it suggests that weighting unigrams by their inverse document frequency may create superior models.

The example above also supports an idea from the literature that suggests that named entities may be more useful predictors of difficulty than common nouns. "Caribbean," "Reyes," and "Sosa" all arguably provide more information, or at least more specific information, than simply "nation" or "players." The specificity of this information also means that there is less chance it appears ambiguously in the dataset; "Sosa" always refers to Sammy Sosa, and Caribbean always refers to a particular region of the Atlantic, though Reyes may also reference different named entities. While there appears to be no pattern to how the model currently uses named entities in its predictions, incorporating information about named entities along with inverse document frequency weights may increase the performance of the model.

Unlike unigram features, features associated with the length of the clue had a significant impact on accuracy. However, this relationship is not as simple as a direct relationship between length and clue difficulty, because the length of a clue can have one of two meanings. In one interpretation, a long clue can indicate a clue that is complex or difficult to process, and a short clue can indicate something that is easy to process. The

model built using only the set of length features correctly predicted the difficulty levels of the following clues, for example:

> MOVIE TITLE OBJECTS $400: “Soylent Green” (people (or food))
> WORLD OF GESTURES $1000: In much of the world, the twirled finger means "you're crazy", but in Argentina, it means you may have one of these--pick up! (phone call)

These clues, which are considered very easy and very difficult, respectively, contrast in the density of the information they provide. In contrast to that example, a long clue could indicate that a contestant has been given multiple pieces of information, which makes giving a correct response easier than if the contestant has only been given one piece of information. However, this could also signal that the clue is perceived as being easier by those assigning labels, since a clue that requires one piece of information to figure out and therefore be labeled very easy. This duality is reflected by the confusion matrix of the model that uses only length features (included in Appendix F), which classifies 18,098 instances (83%) as being either very easy or very difficult. One example of where this conclusion plays out is in the classification of the following “very easy” clues, both of which the model accurately predicted to be “very easy”:

> CROSS WORLD CLUES “C” $200: Sea known as the American Mediterranean (9) (the Caribbean)
> LITERARY E-MAIL ADDRESSES $400: hogwarts_ headmaster@ harrypotter.edu (Dumbledore)

These clues are both relatively short in length going by the number of characters, syllables, and words.

However, the success of length features is limited because many *Jeopardy!* categories pattern their clues similarly across difficulty levels. For example, the category WHAT IS IT YOU DO? has five clues which all consist of one word:

> $200: Thespian (correctly predicted “very easy”)
> $400: Lexicographer (predicted “very easy”)
> $600: Philatelist (predicted “easy”)
> $800: Ichthyologist (correctly predicted “difficult”)
> $1000: Terpsichorean (predicted “very easy”)

The model using only features based on length was able to correctly classify two of the clues, but incorrectly classified the most difficult clue as being very easy. A similar problem occurs in BEFORE & AFTER clues, which require contestants to come up with two different responses and combine them, which involves long clues no matter what the difficulty value. Of the 25 clues in the BEFORE & AFTER category, 8 were classified as being very difficult and 15 were classified as being very easy. The presence of multiple pieces of information in these long clues exacerbates the model’s inability to distinguish between levels of difficulty.

The component phrases of a clue perhaps are closest to capturing the mixture of information, as different noun phrases may contribute different pieces of information to the clue. The number of these phrases, as well as their average length and how many parts the clue was chunked into overall, can indicate whether a clue provides many pieces of information about a topic that are all rather shallow, or whether it provides fewer pieces of more detailed information. The following clues indicate this phenomenon:

> INTERNATIONAL SEAFOOD $200: A Maltese specialty, lampuki pie is made with the fish better known by this double talk name (mahi mahi)
> OPERA $200: "Twilight of the Gods" is the last part of this composer's "Ring" cycle (Wagner)
> “C” IN SCIENCE $1000: The two streams of water form one, because water molecules are so attracted to each other that when they come near, they stick together, a uniting action known as this (cohesion)
> NBA CHAMPS $1000: They must like the odds: behind Tim Duncan, they won in 2003, 2005 & 2007 (the San Antonio Spurs)

In the two clues correctly identified as “very easy,” the number of noun chunks vary but the length is usually at least two words: “A Maltese specialty,” “lampuki pie,” “this double talk name,” “Twilight of the Gods” (parsed as a proper noun), “this composer’s ‘Ring’ cycle.” In contrast, the noun phrases in the two more difficult clues - even though there are more of them for the “C” IN SCIENCE clue - are shorter and signal that less information has been given in a clue. Further research might also explore the impact of phrases containing determiners, such as “this” or “these,” since they identify what entity the clue is quizzing contestants on. Longer phrases containing determiners might signal more information being given about that entity, as opposed to just “this” and “they” in the more difficult clues referenced above.

Media was also a significant predictor of difficulty, though this was more surprising, since there were only three media features, each of which had a small range of values. Within these parameters, there is even less variety than one might expect, as 19,881 clues (89.6%) have all of their media features valued at 0. This distribution of features probably accounts for the fact that a model built using only media features only predicts three of the available five classes, not assigning any clue “easy” or “difficult.” 20,003 (91.6%) clues are predicted to be “very easy” (the majority class) with 1,406 (6.4%) clues assigned “very difficult” and 426 (2.0%) being assigned to the “medium” difficulty class. This reflects some underlying (but likely not meaningful) patterns in the dataset: media is more often found in very difficult and medium difficult clues than it is in any other class (accounting for 475 and 435 clues, respectively, out of 1,954 clues which contain media). In addition, media is least often found in very easy clues, occupying only 273 very easy clues.

Though media is a more extreme version of this phenomenon, each model built using individual sets of readability features had higher recall for the very easy and very difficult clues at the expense of recall for the three intermediate levels. (Confusion matrices that illustrate this point can be found in Appendix F.) While the models usually did not agree on their predictions, they tended to be correct and agree most on short clues without media. For example, the models built using only media, length, and phrase features correctly predicted the following clues to be very easy:

> FILM FIRST NAMES $400: Title spy guy from "The Bourne Identity" (Jason)
> LITERARY BROTHERS $400: Dmitry & Ivan are 2 of this novel's title quartet (*The Brothers Karamazov*)

This might be because these clues have less information in across all three significant categories: they contain no audiovisual content (media), they are relatively short (length), and they have fewer noun phrases that are still relatively information rich, like “Title spy guy” and “this novel’s title quartet” (phrase).

Though media, length and phrase features were somewhat predictive of difficulty, topic was not significantly predictive. Since topic was operationalized as topic membership, it may have had too close a correspondence with the clue categories themselves to be useful in predicting difficulty. Since every category contains one clue from each difficulty level, if the clues’ category corresponded to their topic assignments, these assignments would do little to differentiate between the clues. Even in more miscellaneous categories, such as “POTPOURRI,” topic assignments might be less useful if they were made using less relevant information in the clue. The following medium-difficulty clue was assigned to the “books” topic:

> POTPOURRI $600: Schopenhauer wrote that to do this "is to halve your rights and double your duties"; so he never did (to marry)

However, it could also plausibly be assigned to the "philosophy" topic, or even no topic at all. Schopenhauer's authorship, as opposed to his status as a philosopher, his personal life, or the actual substance of the quote, was used to assign the clue to the books topic, but it may not have been the most important part of the clue. This likely also explains why including synonyms or hypernyms did not significantly increase accuracy; they may have only entrenched the model's existing distortion of what a clue was about.

Topic membership also does not cover the granularity of the information mentioned. Because membership was treated as a binary feature for each topic, the obscurity of the information related to the topic was not included. Schopenhauer, for example, is likely a more obscure author than Charles Dickens, but they would be considered equal members of the literature topic. Dickens might have less overlap with other topics, but a clue's interdisciplinary nature or lack thereof is not a reliable estimate of its difficulty. In addition, using only category names as ways to generate topics meant the topics' granularity was restricted by the show's representation of topics and not any real world representations (for example, "Africa" was a topic, but North America was represented by separate topics for the United States and Canada). Using an external, possibly hierarchical resource to generate topics would create a more objective and possibly more useful method of predicting topics.

Though the topic features did not have a statistically significant impact on accuracy, including any kind of topic features - $topics_{ORIG}$, $topics_{HYP}$ or $topics_{SYN}$ - increased the recall for the medium difficulty class. As with the sets of readability features , all models built at this stage had higher recall for the very easy and very difficult classes, but including topic features prompts the model to guess the medium

difficulty class more often, with a slight improvement in precision for that class as well: 20.1% for $topics_{HYP}$, 20.4% for $topics_{ORIG}$, and 20.4% for $topics_{SYN}$, as opposed to 19.7% precision for the medium difficulty class with only the significant readability features. This might indicate some underlying pattern to the medium difficulty class - perhaps clues at that level are more weighted to certain topics, or perhaps they are more or less likely to be members of multiple topics. However, the number of correct predictions for very easy and very difficult classes is still higher without the topic features, so having higher precision and recall for the medium difficulty class is a trade-off with other predictions.

While the limitations for topic and the earlier ones for unigrams discussed above may have contributed to the models' low accuracy, the significant impact of length, phrase and media features offers a new framework in which to cast these features. Instead of examining unigrams in the context of readability, it will likely be more helpful to consider them in terms of how much information they provide. Disambiguating words by looking particularly at named entities or by considering bigrams or trigrams along with unigrams would create a better representation of how specific the clues are, which might be a better predictor of difficulty. Similarly, detecting more granular topics would allow a model to differentiate between how on-topic clues were.

Though not discussed extensively here, further research might also focus on a different way of approaching the dataset altogether. One assumption made in this study is the independence of the difficulty classes, but as mentioned above, there may be a mixture of easy and difficult information within a clue, which would suggest that not only are the classes on a continuous spectrum, clues in different classes have

commonalities between them. Using difficult clues to predict very difficult clues or using easy clues to predict very easy clues might yield better results than treating the classes as discrete and unrelated.

Conclusion

Though unigrams and topic membership were not shown to be significant features for predicting the difficulty of trivia clues, features relating to media, length and phrases all had significant impact on difficulty. This may be because these features correspond to the amount of information offered in a clue. In contrast, unigrams and topic membership are closer approximations to the subject of a clue than the density of the information it contains. A more concrete representation of the volume of information in a clue would likely prove useful when conducting further research.

Since the significant results are based on the structure of the clues and not their contents, they may generalize to more specific domains, as long as the underlying structure remains the same. Though topics themselves were not significant features for predicting difficulty, any one of the 65 topics used might be used as a domain to test whether the results from this study hold true. In addition, researching clue difficulty only within a specific domain would prevent the model from comparing apples to oranges, which might yield better results for predictions made using unigrams.

This research has also shown how even texts written under constraints have meaningful variations that can be used to infer meaning. Though every *Jeopardy!* clue is rather short, and efforts to expand the text with synonyms and hypernyms were unsuccessful for the purposes of this research, significance still exists within the boundaries of design for each clue. This finding may be useful for those studying Tweets or other documents with constricted forms. Similarly, this research has shown that

knowing the form of media materials linked to by a text is significant without knowing what the media itself contains, suggesting that for similar projects in text mining, gathering or creating exhaustive descriptions of peripheral media files for similar projects may not be necessary.

Among these similar projects might be the application of these results to community question answering systems. This research has attempted to predict the difficulty of a clue without knowledge of the person to whom it is posed, but question recommendations within community question answering systems are usually based on knowledge of the users. In addition, community question answering systems do not usually place restrictions on the cognitive processes required to answer a question, in contrast to the emphasis on remembering information in trivia clues. The applicability of results from this study to other contexts will depend on how trivia-specific they are, and if they can be used to evaluate the difficulty of questions that require higher cognitive processes. While those ideas warrant their own consideration, this research shows that the textual features of a clue or question compared to those of other existing questions can potentially allow question answering systems to measure a question's objective difficulty.

Appendix A

Clues Removed Based on Cognitive Level

MATH-POURRI $800: It's the square root of the square root of 81 (3)
MATH-POURRI $2,000: It's -5 x -5 x -5 (-125)
STORY PROBLEMS $400: 12 hens each lay a half dozen eggs, but 2 get broken; then a hen lays one more, leaving this many unbroken eggs (71)
STORY PROBLEMS $800: If a lepidopterist spots 4 butterflies, each with 3 black spots & 5 yellow spots, he spots this many spots (32)
STORY PROBLEMS $1,200 The "Jeopardy!" writers drank 72 beers after work yesterday: 20 in the 1st round, 26 in the 2nd & this many at last call (26)
STORY PROBLEMS $1,600: A widget costs 64 cents & you pay with a dollar; you get 4 coins in change, these (a penny, a quarter & two nickels)
STORY PROBLEMS $2,000: A scout troop hikes 3 miles due north, 4 miles due east & straight back this far to their starting point (5)
MULTIPLY BY THE CLUE'S VALUE $200: 7 (1400)
MULTIPLY BY THE CLUE'S VALUE $400: 30 (12,000)
MULTIPLY BY THE CLUE'S VALUE $600: 700 (420,000)
MULTIPLY BY THE CLUE'S VALUE $800: 0.6 (480)
MULTIPLY BY THE CLUE'S VALUE $1,000: 10,001 (10,001,000)
x 2 $400: In a bowling game, it's a perfect score x 2 (600)
x 2 $800: It's the total number of sides on an octagon & a nonagon x 2 (34)
x 2 $1,200: It's the number of stars & stripes on the U.S. flag x 2 (126)
TIME $200: 180 minutes after noon (3:00 PM)
TIME $400: 360 seconds before 2:00 am (1:54 AM)
TIME $600: 13 half hours after 6:30 PM (1:00 AM)
TIME $800: 4 dozen minutes after 3:15 PM (4:03 PM)
TIME $1,000: 1,800 seconds before 8:11 AM (7:41 AM)
ALL THINGS CONSIDERED $400: It's 4 cubed minus 4 squared (48)
MULTIPLY BY 5, DIVIDE BY 2, ADD 3 $400: 2 (8)
MULTIPLY BY 5, DIVIDE BY 2, ADD 3 $800: 10 (28)
MULTIPLY BY 5, DIVIDE BY 2, ADD 3 $1,200: 16 (43)
MULTIPLY BY 5, DIVIDE BY 2, ADD 3 $1,600: 60 (153)
MULTIPLY BY 5, DIVIDE BY 2, ADD 3 $2,000: 50,000 (125,003)
GIMME 5 $400: x 16 (80)

GIMME 5 $800: Cubed (25)
GIMME 5 $1,200: x the equivalent of the Roman numeral XX (100)
GIMME 5 $1,600: x the number of degrees in a circle (1800)
GIMME 5 $2,000: x the square root of 121 (55)

Appendix B

Regular Expressions Used to Parse Clues for Phrase Features

After the clues were tagged with their parts of speech, the following regular expressions were used to chunk noun phrases (NP) and verb phrases (VP). The abbreviations for parts of speech are the same used in the Penn Treebank.[7]

Noun phrases:

```
{<DT><JJ><,><JJ><NN>}
{<DT><VBD><[N].*>}
{<DT><POS><JJ><[N.*>}
{<NNP><DT><JJ>}
{<DT>*<NNP>+<\.><NN|NNP>+}
{<``><FW><NNP>+<">}
{<DT><NNS><POS><JJ><VBD><NN>+}
{<DT><VBG><[N].*>*}
{<CD><DT><NNS>}
{<DT><NNP><CD><NN>}
{<DT><CD><NN><CD>}
{<DT>*<NNP><DT|IN><NNP>}
{<DT>*<CD>*<NNP><DT|IN><NNP>+}
{<DT>*<NNP><CD><,><CD><NN>}
{<DT|PRP|CD>}
{<DT>*<[N].*>*<POS><[J].*|VBD>*<[N].*>+<">*<DT>}
{<``><[J].*><NN><">}
{<DT>*<PRP\$>*<RB>*<[J].*><[N].*>+<"><[N].*>*}
{<DT>*<PRP\$>*<RB>*<[J].*><[N].*>+<">*}
{<``><NNP>+<\.>*<">}
{<DT>*<\$|[J].*|RB><CD><[N].*>}
{<DT>*<[J].*><[N].*><CD>}
{<CD><POS>*<[N].*>+<">*}
```

[7] https://www.cis.upenn.edu/~treebank/

{<NNP><">*<POS>*<NNP>*<DT><[J].*><NN|NNP>+<">*}
{<DT><JJ>}

Verb phrases:
{<MD><RB>*<VB><VBN>*}
{<TO><MD>}
{<[V].*>*<RB>*<[V].*>+<RB>*}
{<RB>*<[V].*>*<TO><[V].*>+<RB>*}

Appendix C

List of Category Titles' Unigrams, Bigrams and Trigrams Used to Create Topics

Africa: africa, african
Anatomy: doctor, eyes, health, human_body, medical, medicinal, medicine
Animals: animal, animals, bee, bird, birds, cats, creatures, dog, dogs, fish, mammal
Annual events: annual events, christmas, holidays, observances
Art: art, arte, artist, artists, arts, arty, paintings
Asia: asia, asian
Astronomy: astronomers, astronomical, astronomy, planet
Bible: bible, biblical, testament
Books & literature: authors, bestsellers, book, book_title, books, bookstore, fiction, haiku, literary, literary_characters, literature, nonfiction, novel, page, pages, poems, poet, poetic, poetry, poets, pulitzer, shakespeare, writers
Branding: brand, brand_name, branded, brands, logos, slogan, slogans
Business: business, business_industry, company, organizational
Canada: canada, canadian, canadians
Celebrities: actors, celebrities, celebrity, celebs, hollywood
China: china, chinese
Classical music: composers, opera, operas
College: college, colleges_universities, school, university
Cities & towns: cities, city, town
Dance: ballet, ballets, dance, dances, dancing
English/UK: british, english, england
Fashion: fashion, fashionable
Flags: flag, flags
Food & drink: beer, cuisine, drink, eateries, food, food_drink, foodie, fruits, seafood, soup
Foreign languages: espanol, language, languages
Games: game, games
Geography (all): bodies_water, capital, capitals, city, countries, country, geographic, geography, island, islands, lakes, land, landmark, lands, map, mountains, nation, places, river, rivers, seas, state, states, travel, travels, world
Geography (physical): bodies_water, island, islands, lake, mountains, river, rivers, seas
Geography (political): capital, countries, country, lands, nation, nations, place_names

Germany: german, germany
Gods: god, gods, myths
Greece: greece, greek
History: 80s, 90s, 1600s, 1800s, 1890s, 1930s, 1960s, 1980s, 19th_century, 20th_century, century, historical, history, war, warcraft, wars, year, years
House & home: address, estate, home, house, housekeeping
Inventions: inventions, inventive, inventors
Italy: italian, italiano, italy
Jobs: jobs, work
Leaders: leaders, king, kings, president, presidential, presidents, queen
Military history: battle, battles, civil_war, military, war, wars, warcraft
Money: money
Movies: movies, cinema, movie_characters, films, film, best_picture, screen, movie, movie_title, movie_titles, who_played_em, movie_taglines, moving_pictures
Museum: museum, museums
Music: album, albums, band, bands, beatles, grammy, grammys, ipod, lyrics, music, musical, musicals, no_1_hitmakers, record, singer, song, songs, tunes
Nature: flower, flowers, nature, tree
Opera: opera, operas
Philosophy: philosophy, philosophic
Places: landmark, place, places
Poetry: haiku, poetic, poetry, poets
Politics: cabinet, congressmen, government, governors, leaders, political, politically, politicians, politics, president, presidential, presidents, senator, supreme_court
Presidents: president, presidential, presidents
Publications: magazines, newspapers
Religion: bible, biblical, pope, religions, religious, saints, testament
Royalty: king, kings, queen
Russia: russia, russian
Science: biology, element, elemental, elements, physics, science, scientists
Shakespeare: shakespeare, shakespearean
Spain: spain, spanish
Sports: athletics, ballpark, baseball, football, nba, nfl, racing, sporting, sports, sports_illustrated, sports_teams, sportsmen
Technology: computer, internet, online, website, websites
Theater: broadway, drama, dramas, musicals, play_title, shakespeare, theater
TV: episodes, TV, TV_series,
US (country): america, american, americana, americans, president, presidential, presidents, u_s
US (states): state, states, official_state

Weather: weather
Weights & measures: measure, weights_measures
Word origins: etymology, from_french, from_german, from_greek, word_origins, words_from,
World: world, international

Appendix D

Accuracy and F-measures of Topic Models

Topic	Clues in training set	Clues in test set	Accuracy (%)	F-measure (topic)	F-measure (non-topic)
Africa	94	25	0.96	0.965517241	0.952380952
anatomy	373	95	0.9263	0.929292929	0.923076923
animals	855	215	0.8558	0.870292887	0.837696335
annual events	125	33	0.8788	0.894736842	0.857142857
art	665	169	0.9349	0.937142857	0.932515337
Asia	83	23	0.913	0.923076923	0.9
astronomy	117	31	0.9355	0.9375	0.933333333
Bible	285	73	0.9452	0.948717949	0.941176471
books & literature	2621	783	0.931	0.925	0.936170213
branding	231	59	0.8814	0.895522388	0.862745098
business	285	73	0.9452	0.945945946	0.944444444
Canada	119	31	0.8387	0.864864865	0.8
celebrities	389	99	0.9192	0.921568627	0.916666667
China	71	19	0.7368	0.761904762	0.705882353
cities & towns	403	103	0.8738	0.888888889	0.853932584
classical music	359	91	0.9121	0.92	0.902439024
college	215	55	0.8545	0.875	0.826086957
dance	203	53	0.9811	0.981132075	0.981132075
English/UK	297	77	0.8701	0.875	0.864864865
fashion	100	27	0.8519	0.846153846	0.857142857
flags	95	25	1	1	1
food & drink	751	189	0.9683	0.969072165	0.967391304
foreign language	175	45	0.7111	0.779661017	0.580645161
games	163	43	0.907	0.916666667	0.894736842
geography (all)	3013	755	0.8728	0.878172589	0.867036011
geography (physical)	409	103	0.8835	0.896551724	0.866666667

geography (political)	865	219	0.8356	0.853658537	0.8125
German	63	17	0.8824	0.909090909	0.833333333
gods & myths	191	49	0.9184	0.925925926	0.909090909
Greece	149	39	0.8718	0.888888889	0.848484848
history	2123	533	0.8818	0.880455408	0.883116883
house & home	131	33	0.7575	0.692307692	0.8
inventions	71	19	0.8421	0.857142857	0.823529412
Italy	65	19	0.7368	0.8	0.615384615
jobs	53	15	0.6	0.5	0.666666667
leaders	521	133	0.8722	0.88590604	0.854700855
military history	632	127	0.9921	0.992248062	0.992
money	55	15	1	1	1
movies	1754	440	0.9227	0.926724138	0.918269231
museum	125	33	0.9697	0.971428571	0.967741935
music	1880	377	0.9788	0.979274611	0.97826087
nature	195	51	0.8824	0.892857143	0.869565217
opera	217	55	0.9273	0.933333333	0.92
philosophy	55	15	0.8667	0.888888889	0.833333333
places	127	33	0.8485	0.87804878	0.8
poetry	297	77	0.9351	0.939759036	0.929577465
politics	745	189	0.9312	0.935960591	0.925714286
presidents	349	89	0.9213	0.927835052	0.913580247
publications	113	17	1	1	1
religion	469	119	0.9244	0.930232558	0.917431193
royalty	185	49	0.898	0.901960784	0.893617021
Russia	100	21	1	1	1
science	849	215	0.9163	0.920353982	0.911764706
Shakespeare	303	77	0.9091	0.917647059	0.898550725
Spain	71	19	0.7895	0.833333333	0.714285714
sports	849	213	0.9671	0.968325792	0.965853659
technology	141	37	0.9459	0.947368421	0.944444444
theater	695	175	0.9371	0.941176471	0.932515337
TV	757	191	0.9424	0.945273632	0.939226519
US (country)	1269	319	0.8809	0.886904762	0.874172185
US (states)	631	161	0.8509	0.870967742	0.823529412
weather	61	17	0.9412	0.947368421	0.933333333
weights &	95	25	0.8	0.838709677	0.736842105

measures					
word origins	253	65	0.9077	0.914285714	0.9
world	922	232	0.9181	0.92244898	0.913242009

Appendix E

Features and Their Values

Feature group	Feature set	Component features	Value type (range, if applicable)
Readability	Unigrams	unigrams	binary
	Media	number of pictures	integer (0-4)
		number of audio files	integer (0-2)
		number of video files	integer (0-2)
	Length	number of characters	integer (1-342)
		number of syllables	integer (0-87)
		number of words	integer (0-62)
		number of numbers	integer (0-10)
		number of words and numbers	integer (0-62)
		characters per word	decimal (1-15)
		syllables per word	decimal (1-6)
	Phrase	number of noun phrases	integer (0-20)
		number of verb phrases	integer (0-11)
		average length in words of noun phrases	decimal (0-8)
		average length in words of verb phrases	decimal (0-6)
		total number of phrases (noun, verb, and other)	integer (1-61)
Topic	Topic	membership for 65 topics	binary
		total number of topics clue belongs to	integer (0-65)

Appendix F

Confusion Matrices

unigrams$_{10}$

predicted \ **actual**	*very easy*	*easy*	*medium*	*difficult*	*very difficult*
very easy	1234	1221	1211	1164	1070
easy	914	830	780	753	726
medium	762	781	728	749	753
difficult	775	789	820	808	825
very difficult	731	772	845	876	918

Table F1. Confusion matrix for a model built using only the unigrams$_{10}$ set of features.

length

predicted \ **actual**	*very easy*	*easy*	*medium*	*difficult*	*very difficult*
very easy	2129	1945	1872	1821	1699
easy	362	339	338	301	315
medium	217	200	231	213	213
difficult	190	220	200	191	217
very difficult	1518	1689	1743	1824	1858

Table F2. Confusion matrix for a model built using only the length set of features.

phrase

predicted \ **actual**	*very easy*	*easy*	*medium*	*difficult*	*very difficult*
very easy	2115	1981	1910	1906	1789
easy	853	949	904	858	932
medium	132	145	145	155	177
difficult	410	403	408	435	429
very difficult	906	915	1017	996	965

Table F3. Confusion matrix for a model built using only the phrase set of features.

media		**actual**				
		very easy	*easy*	*medium*	*difficult*	*very difficult*
predicted	*very easy*	4164	4018	3975	4006	3840
	easy	0	0	0	0	0
	medium	68	81	92	80	105
	difficult	0	0	0	0	0
	very difficult	184	294	317	264	347

Table F4. Confusion matrix built for a model using only the media set of features.

readability (media, length, phrase)		**actual**				
		very easy	*easy*	*medium*	*difficult*	*very difficult*
predicted	*very easy*	2249	2040	1949	1920	1769
	easy	341	373	377	328	345
	medium	174	180	181	193	196
	difficult	523	590	552	569	591
	very difficult	1129	1210	1325	1340	1391

Table F5. Confusion matrix for a model built using only significant readability features.

topics$_{ORIG}$		**actual**				
		very easy	*easy*	*medium*	*difficult*	*very difficult*
predicted	*very easy*	2027	1858	1778	1704	1622
	easy	351	343	369	351	323
	medium	700	717	731	744	685
	difficult	472	513	534	526	541
	very difficult	866	962	972	1025	1121

Table F6. Confusion matrix for a model built using only the topics$_{ORIG}$ set of features.

topics$_{HYP}$		**actual**				
		very easy	*easy*	*medium*	*difficult*	*very difficult*
predicted	*very easy*	1798	1673	1670	1626	1568
	easy	296	260	286	286	256
	medium	1021	1060	1044	1006	945
	difficult	527	532	544	540	546

	very difficult	774	868	840	892	977

Table F7. Confusion matrix for a model built using only the $topics_{HYP}$ set of features.

$topics_{SYN}$		**actual**				
		very easy	*easy*	*medium*	*difficult*	*very difficult*
	very easy	1994	1769	1754	1646	1542
	easy	461	460	440	461	453
predicted	*medium*	662	635	672	679	645
	difficult	501	578	560	542	593
	very difficult	798	951	958	1022	1059

Table F8. Confusion matrix for a model built using only the $topics_{SYN}$ set of features.

References

Aksoy, C., Can, F., & Kocberber, S. (2012). Novelty detection for topic tracking. *Journal of the American Society for Information Science and Technology, 63*(4), 777-795.

Bunescu, R., & Huang, Y. (2010, October). Learning the relative usefulness of questions in community QA. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 97-107). Association for Computational Linguistics.

Burel, G., Mulholland, P., He, Y., & Alani, H. (2015, August). Predicting Answering Behaviour in Online Question Answering Communities. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 201-210). ACM.

Cai, L., Zhou, G., Liu, K., & Zhao, J. (2011). Learning the Latent Topics for Question Retrieval in Community QA. In *IJCNLP* (Vol. 11, pp. 273-281).

Carnegie Mellon University. (2015). LightSide (Version 2.3.2) [Software]. Available from http://ankara.lti.cs.cmu.edu/side/

Carroll, A. R. (2015). Exploring The Effects Of Multimedia Content On A Question And Answer System. Retrieved from All Theses. Paper 2084.

Collins‐Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology, 56*(13), 1448-1462.

Diao, H., Xu, G., & Xiao, J. (2011). An improved new event detection model. In *Information and Automation* (pp. 431-437). Springer Berlin Heidelberg.

Dror, G., Koren, Y., Maarek, Y., & Szpektor, I. (2011, August). I want to answer; who has a question?: Yahoo! answers recommender system. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1109-1117). ACM.

Duan, H., Cao, Y., Lin, C. Y., & Yu, Y. (2008, June). Searching Questions by Identifying Question Topic and Question Focus. In *ACL* (pp. 156-164).

Eyecioglu, A., & Keller, B. (2015). ASOBEK: Twitter paraphrase identification with simple overlap features and SVMs. Proceedings of *SemEval*.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010, August). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 276-284). Association for Computational Linguistics.

Gaffney, M. (2006). Gridlock: crossword puzzles and the mad geniuses who create them. Thunder's Mouth Press.

Heilman, M. (2011). Automatic factual question generation from text (Doctoral dissertation, Carnegie Mellon University). Retrieved from http://errico.srv.cs.cmu.edu/research/thesis/2011/michael_heilman.pdf

J! Archive. (2014). Jeopardy! [data file]. Retrieved from https://drive.google.com/file/d/0BwT5wj_P7BKXUl9tOUJWYzVvUjA/view

Jennings, K. (2007). Brainiac: adventures in the curious, competitive, compulsive world of trivia buffs. Villard Books.

Ji, Z., Xu, F., Wang, B., & He, B. (2012, October). Question-answer topic model for question retrieval in community question answering. In *Proceedings of the 21st*

ACM international conference on Information and knowledge management (pp. 2471-2474). ACM.

Karampatsis, R. M. (2015). CDTDS: Predicting paraphrases in Twitter via support vector regression. *Proceedings of SemEval*.

Karkali, M., Rousseau, F., Ntoulas, A., & Vazirgiannis, M. (2013). Efficient online novelty detection in news streams. In *Web Information Systems Engineering–WISE 2013* (pp. 57-71). Springer Berlin Heidelberg.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212-218.

Kumaran, G., & Allan, J. (2005, October). Using names and topics for new event detection. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 121-128). Association for Computational Linguistics.

Li, S., & Manandhar, S. (2011, June). Improving question recommendation by exploiting information need. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 1425-1434). Association for Computational Linguistics.

Library of Congress. (2016). *Library of Congress classification outline*. Retrieved from https://www.loc.gov/catdir/cpso/lcco/

Liu, X., Croft, W. B., Oh, P., & Hart, D. (2004, July). Automatic recognition of reading levels from user queries. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 548-549). ACM.

Liu, J., Wang, Q., Lin, C. Y., & Hon, H. W. (2013). Question Difficulty Estimation in Community Question Answering Services. In *EMNLP* (pp. 85-90).

Machine Learning Group at the University of Waikato. (2015). Weka (3.6) [Software]. Available from http://www.cs.waikato.ac.nz/ml/weka/index.html

McCown, A. (2015, March 12). What's it like to be one of the Jeopardy! clue writers? Retrieved from http://www.avclub.com/article/whats-it-be-one-jeopardy-clue-writers-216093

Ng, K. W., Tsai, F. S., Chen, L., & Goh, K. C. (2007, December). Novelty detection for text documents using named entity recognition. *In Information, Communications & Signal Processing, 2007 6th International Conference* on (pp. 1-5). IEEE.

Nomoto, T. (2010, October). Two-tier similarity model for story link detection. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 789-798). ACM.

Otterbacher, J., & Radev, D. (2006, August). Fact-focused novelty detection: A feasibility study. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 687-688). ACM.

Pal, A., & Konstan, J. A. (2010, October). Expert identification in community question answering: exploring question selection bias. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1505-1508). ACM.

Paukkeri, M. S., Ollikainen, M., & Honkela, T. (2013). Assessing user-specific difficulty of documents. *Information Processing & Management, 49*(1), 198-212.

Pennington, J., Socher, R., Manning, C.D. (2014). GloVe: Global Vectors for Word Representation. Retrieved from http://nlp.stanford.edu/projects/glove/

Petrović, S., Osborne, M., & Lavrenko, V. (2012, June). Using paraphrases for improving first story detection in news and Twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 338-346). Association for Computational Linguistics.

Princeton University. (2015). *WordNet*. Retrieved from https://wordnet.princeton.edu/

Romano, M. (2005). Crossworld: one man's journey into America's crossword obsession. Broadway Books.

Sanborn, A., & Skryzalin, J. (2015, June 5). Deep Learning for Semantic Similarity. Retrieved from https://cs224d.stanford.edu/reports/SanbornAdrian.pdf

Singh, A. (2012, July). Entity based q&a retrieval. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1266-1277). Association for Computational Linguistics.

Smucker, M. D., Allan, J., & Carterette, B. (2007, November). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 623-632). ACM.

Tanaka-Ishii, K., Tezuka, S., & Terada, H. (2010). Sorting texts by readability. *Computational Linguistics*, *36*(2), 203-227.

Tsai, F. S., & Kwee, A. T. (2011). Experiments in term weighting for novelty mining. *Expert Systems with Applications*, *38*(11), 14094-14101.

University of Pennsylvania. (1999). *The Penn Treebank project*. Retrieved from https://www.cis.upenn.edu/~treebank/

Wanas, N., Magdy, A., & Ashour, H. (2009). Using automatic keyword extraction to detect off-topic posts in online discussion boards. *Proceedings of Content Analysis for Web*, 2.

Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., & Chen, Z. (2013, October). Cqarank: jointly model topics and expertise in community question answering. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 99-108). ACM.

Zhang, Y., & Tsai, F. S. (2009, February). Combining named entities and tags for novel sentence detection. In *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval* (pp. 30-34). ACM.

Zhou, G., Chen, Y., Zeng, D., & Zhao, J. (2013, October). Towards faster and better retrieval models for question search. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (pp. 2139-2148). ACM.

Zhou, G., Lai, S., Liu, K., & Zhao, J. (2012, October). Topic-sensitive probabilistic model for expert finding in question answer communities. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1662-1666). ACM.

Zhou, T. C., Lin, C. Y., King, I., Lyu, M. R., Song, Y. I., & Cao, Y. (2011, April). Learning to Suggest Questions in Online Forums. In *AAAI*.