Question Classification based on LSTMs

Group A:

Michal Harakal, Petra Woßeng, Andreas Neutze, Sven Wontroba

04.12.2020

0. Introduction

Task

Given data on more than 5000 labeled questions of a Question Answering (QA) system, the task is to build a text classifier that can distinguish questions in different categories.

Dataset

The training dataset contains ca. 5500 labelled questions, the testing dataset another 500.

The dataset has six labels, including Abbreviation (ABBR), Entity (ENTY), Description (DESC), Human (HUM), Location (LOC) and Numeric (NUM). More information on the labels in the dataset is presented in [1]. Average length of each sentence is ten, with a vocabulary size of 8700. The training and test txt files can be downloaded from [2], [3],

The dataset is composed of two main attributes (text, label).

[1] https://cogcomp.seas.upenn.edu/Data/QA/QC/definition.html

[2] https://cogcomp.seas.upenn.edu/Data/QA/QC/train_5500.label

[3] https://cogcomp.seas.upenn.edu/Data/QA/QC/TREC_10.label

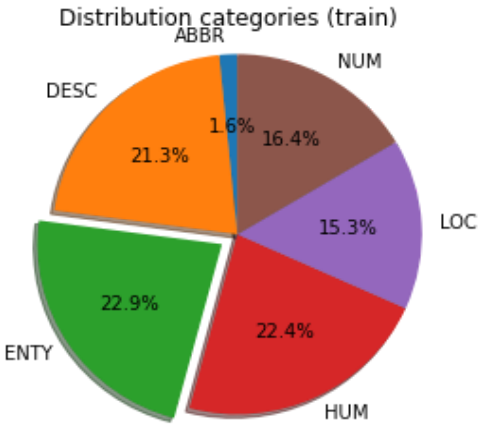
Our Milestones

1. Data Analysis and Preparation
2. Build a LSTM text classifier
3. Use a SVM for the categorization of questions
4. Compare Models incl. Visualizations and scores (F1)
5. Conclusion and Outlook

1. Data Analysis and Preparation

Exploratory Data Analysis using pandas showed 70 duplicates in the training dataset. Those were eliminated and hence, further analysis was based on 5.382 observations with three attributes: question text, category and subcategory.

Most frequent categories are ENTY (entity) and HUM (human), both with more than 1.200 observations; by far the smallest category is ABBR (abbreviation) , which is the label for only 86 questions.



Whereas the 5 main categories are not particularly unbalanced, this is the case for the subcategories. The by far most frequent (of in total nearly 50) subcategories is “HUMan-Individual”, accounting for approx. 20% of all questions’ labels, followed by “DESCription-Definition” with approx. 9%. - Furthermore, in 3 of the 6 main categories, further subcategories are summarized in an “other” subcategory.

Due to the fact that the subcategories are highly unbalanced (and because of the “black-box character” of the three not so small “other” subcategories, we focussed on the main categories.

As the objective of the project was the classification of categories by the questions’ text, some more text analysis and preprocessing was carried out.

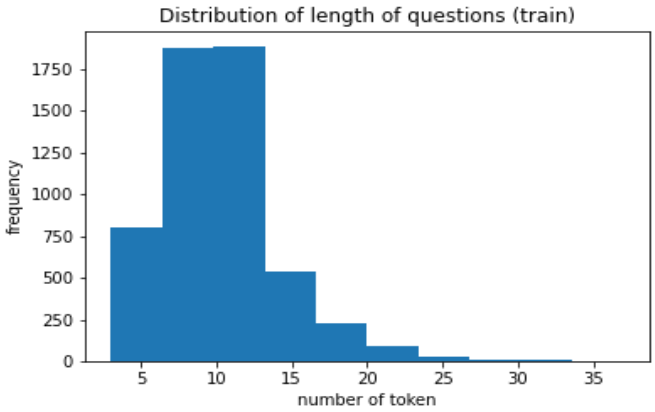
The beginning of each question is not surprisingly dominated by question words. In our dataset “What” was the first word in more than 60% of all cases, followed by How-Who-Where-When-Which-Why.

Example are shown in c) Report the top 10 representative (most repeated) 1, 2 and 3-grams for each of

When referring question words to the associated categories, it was found that

- Questions relating to “ENTITY” were almost always introduced by “What” (1.100 cases)
- In addition, “What” is also highly related to “DESCription” (>700 cases)
- “Who” nearly always leads to the category “HUMAN”
- “How” (many, much) mostly asks for NUMeric values
- …

Another analysis was about the length of the question text, showing a mean of 10 tokens (words, punctuation marks, special characters). The length distribution obviously is not symmetric, about 10% of the questions are of length 15 to 37 tokens.



As common practice in text analysis, some cleanup and shrinkage of the texts was then carried out. To get rid of stopwords, the NLTK package was applied. Final cleaning covered the deletion of question marks and others.

Finally, two features with reduced size of tokens were generated: At first, the question (text) was shortened to a new feature ‘text’ after the deletion of stopwords and in a second step we obtained ‘text_clean’ after the cleaning procedure. This resulted in a considerably shorter text length with a mean of only6 (compared to 10) and a maximum length of 20 (compared to 37).

Note, that in case of questions one should not transform the complete text into lowercase, because otherwise the question words would be regarded as stopwords and then deleted.

2. LSTM

Data preparation

1. Tokenizing
2. Sequences
3. Labels encoding

Tokenizing

As a tokenizer we have used the default tokenizer from **Keras.** Important is and we have learned this a hard way (with a terrible accuracy result 0.26), that in vocabulary for both data sets, the tokens have to have the same indexes.

Sequences

Because we have different max length of question sentences in training and test data, where the maximal length of question in test data is shorter , we have to agree on the same length for both sets. Since we don't want to lose any information from the training set first, the maximal length of the training set is to use for both.

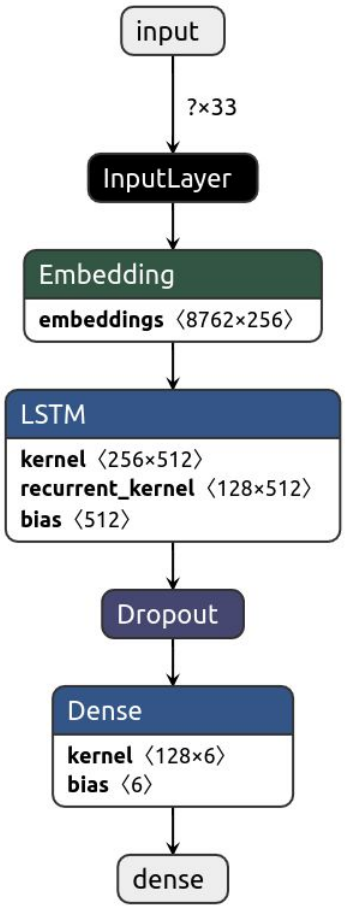
Categories encoding

In this project we have worked with main categories. They were encoded with One Hot encoding with the same values on both data sets.

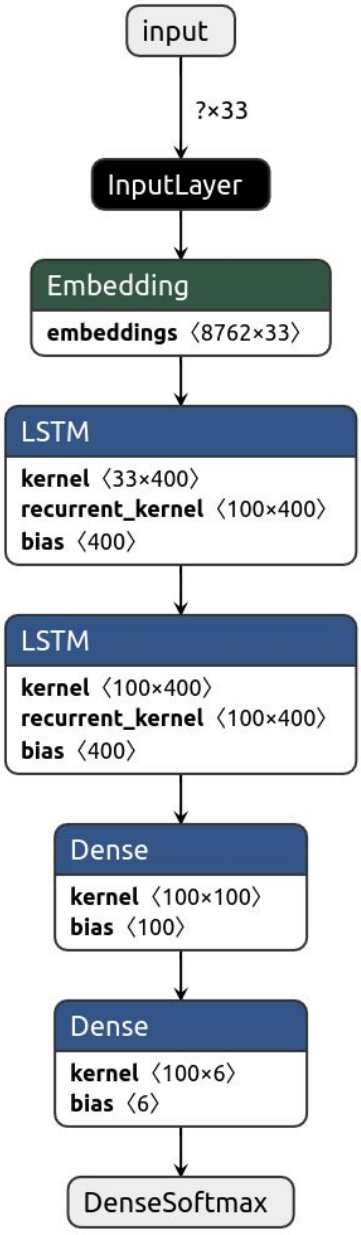
Models

For this project 3 different models have been created. Model creation, training and validation was conducted by usage of Keras library.

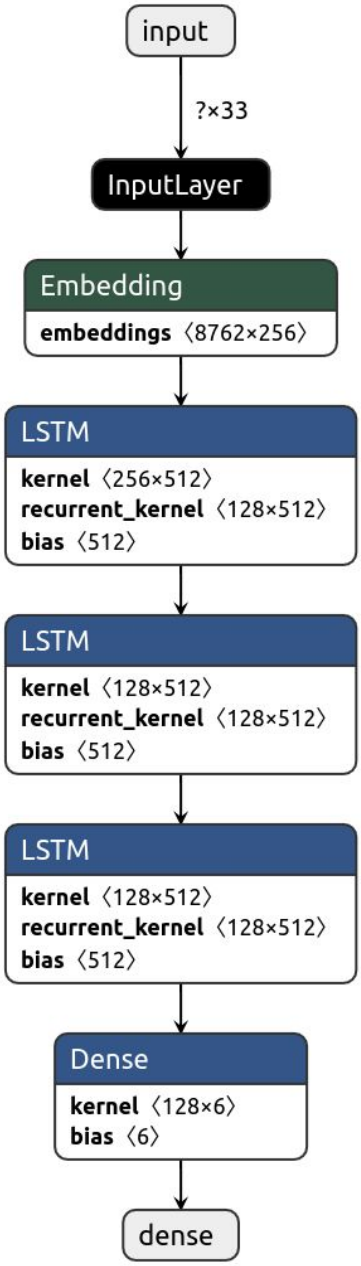
1. Simple LSTM model with dropout. We have find out that a simple model without dropout overfit immediately with very bad validation results on the test set.
2. Because order of word in the sentence is important to classification results, we have also experimented with networking with multiple LSTM layers.



Simple LSTM with dropout



Double LSTM with double Dense

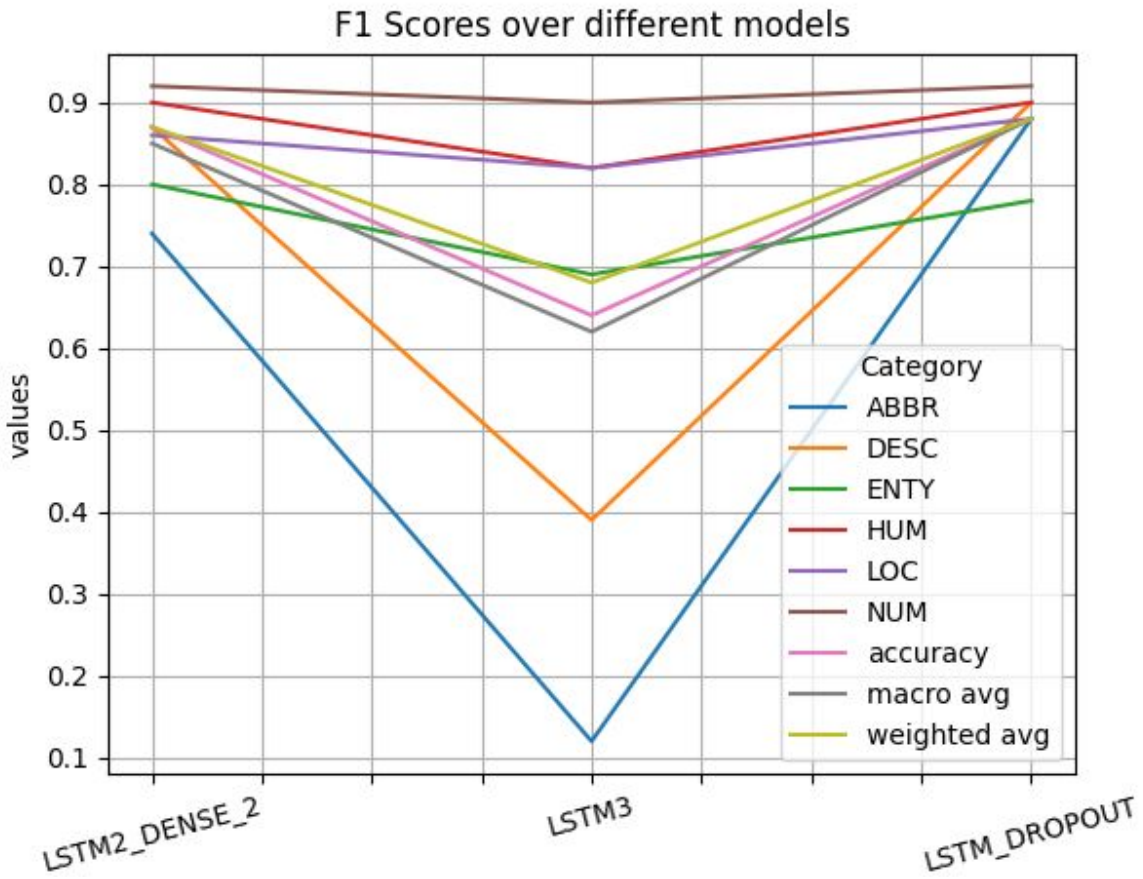


Triple LSTM

Following numbers are based on the data sets (33 is the maximal length of sequence, coming from the longest question). 8762 ist vocabulary size)

LSTM models comparison

As measurement for models comparison was taken weighted average of recall and precision, **F1-score**



LSTM models comparison result

Simple **LSTM model with dropout** provided the best results. This model is also used to comparison with the best SVM model.

3. Use a SVM for the categorization of questions

The following subtasks were to be addressed:

a. Do the same categorization using a Support Vector Machine (SVM) based approach.

b. Test different lengths of N-Grams and report the best N in which the classifier achieved the best performance

c. Report the top 10 representative (most repeated) 1, 2 and 3-grams for each of the classes.

d. Analyze the outcomes.

a) Two initial SVM approaches

Using the svm.LinearSVC() model from sklearn we built different pipelines with CountVectorizer and TfidfVectorizer. In both cases we achieved a cross validation score of about 83% with the base model. Based on the task description we continued with the CountVectorizer to try different n-grams length. With longer n-gram size we got a slightly better overall accuracy result and f1-scores which will be discussed in b) Testing different N-Grams

b) Testing different N-Grams

When specifying n-grams of length 2, 3, ..., it was found that for n-grams = 2 the cross_val_score improved up to 0.86, but for n-grams=3 it slightly decreased.

Differences in scores for different n-gram settings and data inputs generate.

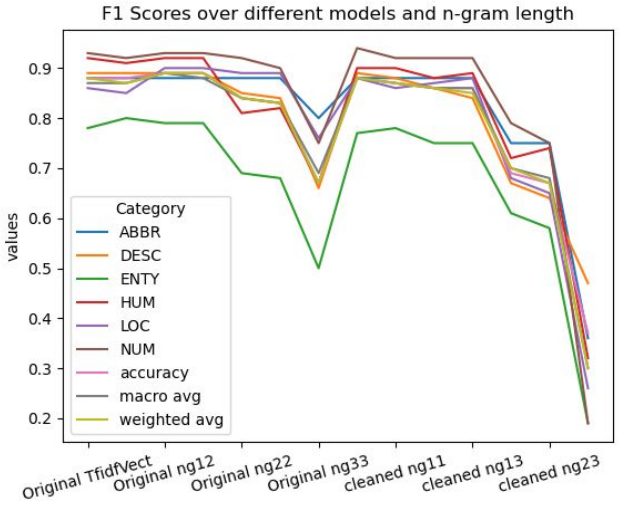
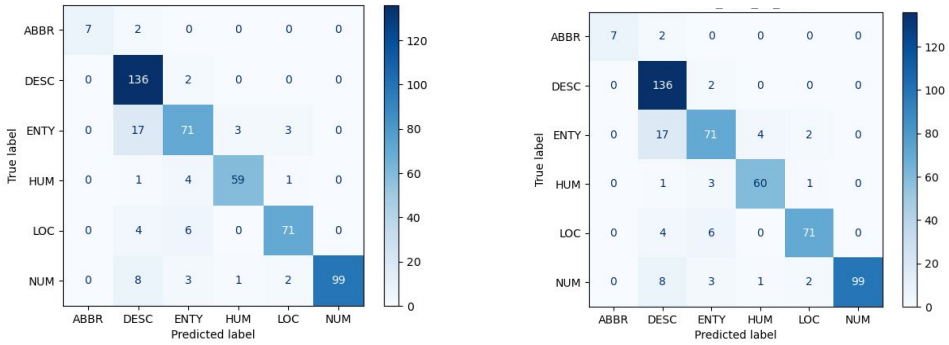


Image: Classification Report for Questions not modified - several n-gram lengths defined in table



Images: Confusion Matrix for Questions not modified default n-gram length of 1 and 1-2

The Confusion Matrix of the cleaned text looks similar so we will not present it here.

Images can be generated with svm_workflow.py and found in sub directory plots

Findings: N-Gram length till 3 allowing smaller like 1 till 3 in the count vectorizer works well. Bigger ones or strict to 2 or 3 and bigger do not work well. Based on values 1-2 works the best.

c) Report the top 10 representative (most repeated) 1, 2 and 3-grams for each of the classes

To show the different and challenges we had with the task we show the Unfiltered Text and later cleaned to compare.

Questions without modification processed just default CountVectorizer settings

stop_words=None

As an example for unigrams see:

ngram_range: (1, 1)

Category: DESC

words: ['what', 'the', 'is', 'how', 'of', 'do', 'in', 'to', 'does', 'are']

(...)

As another example with uni-, bi, and trigrams see:

ngram_range: (1, 3)

Category: DESC

words: ['what', 'the', 'is', 'what is', 'how', 'of', 'do', 'is the', 'what is the', 'in']

Full comparison see appendix

d) Analyze the outcome

Finally, categories were predicted based on the test data set (->F1 scores)

Based on the findings during the experiments we can say a n-gram range of (1,2) creates the best result for all categories. As seen in the confusion matrix for some categories, a setting of (1,3) is slightly better but also degrades the result of other categories thus a setting of (1,2) is recommended.

The results are “visualized” reviewing the n-gram most repeated words. An interesting aspect of this task was to learn that compared to the normal thought workflow - cleaning the text, applying a lemmatizer and removing stopwords in this task does not give us better results.

This seems to be related to the size of questions and structure of them. Here the machine learns and differentiates similar to the human brain and needs the stopword to separate the categories.

Small outlook: To get better results not just in the differentiation of the categories also being able to separate the subcategories it is necessary to develop additional features as input.

Some aspect of this will be described in Conclusion and Outlook

4. Compare Models incl. Visualizations and scores (F1)

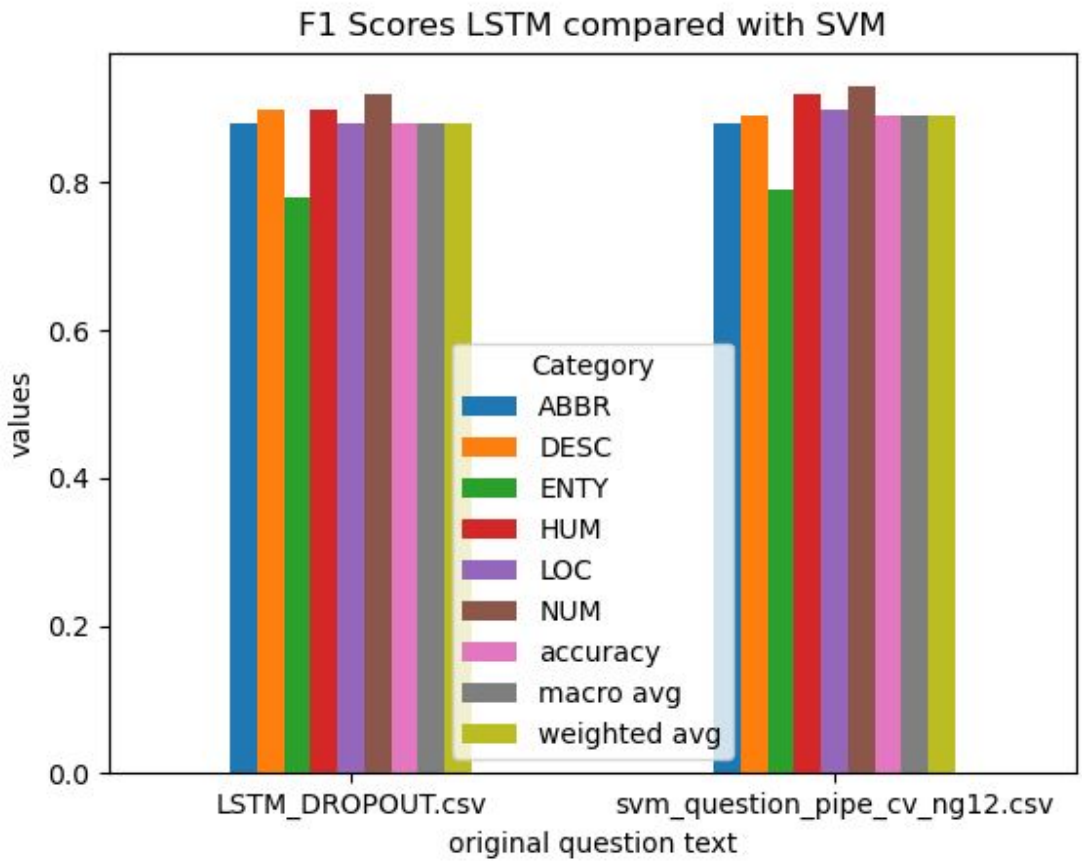


Image: Classification Report for Questions not modified - several n-gram lengths defined in table

Both models are similar in their scores - have only small differences in some of the categories

Both model to compare result processed the original text without cleaning

Interesting findings: The LSTM was better in categories where the SVM had problems with.

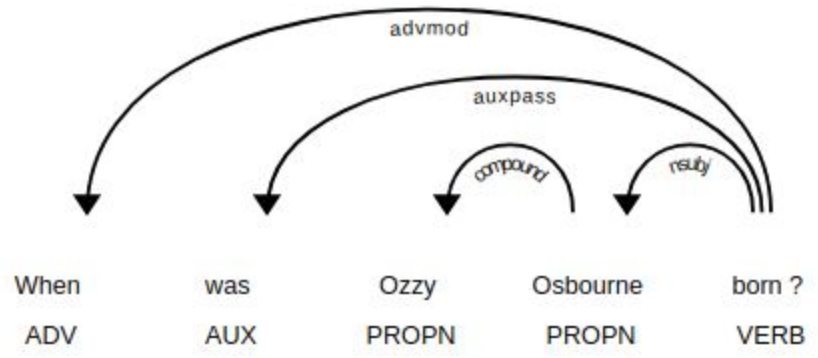
5. Conclusion and outlook

Both models resulted in high accuracies on the testing data, see F1-score.

Further approaches would include a deeper usage of hyperparameter tuning. First variants of our models only showed slight improvements when tunig the regularization parameter C in SVM, e.g.

An idea for further analysis would be the use spaCy in order to include more features of the text, e.g. POS, see the example in our data set below, which is the visualization of the POS tags in one of the first questions:

```
When was Ozzy Osbourne born ?
```



Another possibility to enrich the analysis could be a more detailed investigation of the properties and characteristics of each word/token, e.g. lower case, alphanumeric characters, stop words etc.

```
token - lemma_ - lower_- pos_ - tag_- dep_ - sentiment - is_alpha - is_stop
-------------------------------------------------------------------------
0 When - when - when - ADV - WRB - advmod - 0.0 - True - True
1 was - be - was - AUX - VBD - auxpass - 0.0 - True - True
2 Ozzy - Ozzy - ozzy - PROPN - NNP - compound - 0.0 - True - False
3 Osbourne - Osbourne - osbourne - PROPN - NNP - nsubj - 0.0 - True - False
4 born - bear - born - VERB - VBN - ROOT - 0.0 - True - False
5 ? - ? - ? - PUNCT - . - punct - 0.0 - False - False
```

A. Appendix

['DESC' 'ENTY' 'ABBR' 'HUM' 'NUM' 'LOC']

Text not modified

ngram_range: (1, 1)

Category: DESC words: ['what', 'the', 'is', 'how', 'of', 'do', 'in', 'to', 'does', 'are']

words: wordcount what : 762, the : 603, is : 487, how : 287, of : 247, do : 213, in : 130, to : 130, does : 129, are : 106,

Category: ENTY words: ['what', 'the', 'of', 'is', 'in', 'to', 'was', 'name', 'are', 'for']

words: wordcount what : 1152, the : 876, of : 452, is : 400, in : 302, to : 163, was : 138, name : 119, are : 115, for : 109,

Category: ABBR words: ['what', 'for', 'the', 'does', 'stand', 'is', 'abbreviation', 'of', 'in', 'mean']

words: wordcount what : 86, for : 54, the : 48, does : 48, stand : 41, is : 33, abbreviation : 16, of : 15, in : 10, mean : 9,

Category: HUM words: ['the', 'who', 'what', 'of', 'in', 'was', 'is', 'to', 'name', 'and']

words: wordcount the : 1019, who : 597, what : 547, of : 387, in : 299, was : 291, is : 234, to : 155, name : 144, and : 131,

Category: NUM words: ['the', 'how', 'many', 'what', 'of', 'in', 'is', 'when', 'are', 'did']

words: wordcount the : 560, how : 492, many : 323, what : 276, of : 246, in : 245, is : 223, when : 136, are : 136, did : 121,

Category: LOC words: ['the', 'what', 'is', 'where', 'in', 'of', 'country', 'city', 'can', 'to']

words: wordcount the : 669, what : 554, is : 304, where : 258, in : 221, of : 201, country : 123, city : 101, can : 77, to : 75,

ngram_range: (1, 2)

Category: DESC words: ['what', 'the', 'is', 'what is', 'how', 'of', 'do', 'is the', 'in', 'to']

words: wordcount what : 762, the : 603, is : 487, what is : 420, how : 287, of : 247, do : 213, is the : 194, in : 130, to : 130,

Category: ENTY words: ['what', 'the', 'of', 'is', 'in', 'what is', 'is the', 'to', 'was', 'name']

words: wordcount what : 1152, the : 876, of : 452, is : 400, in : 302, what is : 253, is the : 171, to : 163, was : 138, name : 119,

Category: ABBR words: ['what', 'for', 'the', 'does', 'what does', 'stand', 'stand for', 'is', 'what is', 'is the']

words: wordcount what : 86, for : 54, the : 48, does : 48, what does : 48, stand : 41, stand for : 41, is : 33, what is : 30, is the : 16,

Category: HUM words: ['the', 'who', 'what', 'of', 'in', 'was', 'is', 'of the', 'to', 'name']

words: wordcount the : 1019, who : 597, what : 547, of : 387, in : 299, was : 291, is : 234, of the : 156, to : 155, name : 144,

Category: NUM words: ['the', 'how', 'many', 'how many', 'what', 'of', 'in', 'is', 'when', 'is the']

words: wordcount the : 560, how : 492, many : 323, how many : 323, what : 276, of : 246, in : 245, is : 223, when : 136, is the : 136,

Category: LOC words: ['the', 'what', 'is', 'where', 'in', 'of', 'is the', 'country', 'city', 'what is']

words: wordcount the : 669, what : 554, is : 304, where : 258, in : 221, of : 201, is the : 139, country : 123, city : 101, what is : 95,

ngram_range: (2, 2)

Category: DESC words: ['what is', 'is the', 'how do', 'of the', 'what are', 'what does', 'do you', 'the origin', 'origin of', 'how can']

words: wordcount what is : 420, is the : 194, how do : 125, of the : 85, what are : 77, what does : 68, do you : 64, the origin : 53, origin of : 53, how can : 48,

Category: ENTY words: ['what is', 'is the', 'of the', 'what was', 'in the', 'what are', 'was the', 'what the', 'fear of', 'name of']

words: wordcount what is : 253, is the : 171, of the : 103, what was : 78, in the : 75, what are : 75, was the : 63, what the : 62, fear of : 62, name of : 62,

Category: ABBR words: ['what does', 'stand for', 'what is', 'is the', 'the abbreviation', 'does the', 'abbreviation for', 'does stand', 'abbreviation of', 'of the']

words: wordcount what does : 48, stand for : 41, what is : 30, is the : 16, the abbreviation : 14, does the : 12, abbreviation for : 7, does stand : 5, abbreviation of : 5, of the : 5,

Category: HUM words: ['of the', 'who was', 'was the', 'is the', 'who is', 'in the', 'what is', 'the first', 'name of', 'what was']

words: wordcount of the : 156, who was : 144, was the : 124, is the : 107, who is : 107, in the : 99, what is : 62, the first : 61, name of : 55, what was : 49,

Category: NUM words: ['how many', 'is the', 'what is', 'are there', 'was the', 'in the', 'how much', 'when was', 'how long', 'when did']

words: wordcount how many : 323, is the : 136, what is : 111, are there : 71, was the : 67, in the : 61, how much : 57, when was : 53, how long : 50, when did : 49,

Category: LOC words: ['is the', 'what is', 'what country', 'where is', 'in the', 'where can', 'of the', 'the world', 'what city', 'can find']

words: wordcount is the : 139, what is : 95, what country : 86, where is : 77, in the : 74, where can : 67, of the : 62, the world : 57, what city : 53, can find : 44,

ngram_range: (1, 3)

Category: DESC words: ['what', 'the', 'is', 'what is', 'how', 'of', 'do', 'is the', 'what is the', 'in']

words: wordcount what : 762, the : 603, is : 487, what is : 420, how : 287, of : 247, do : 213, is the : 194, what is the : 179, in : 130,

Category: ENTY words: ['what', 'the', 'of', 'is', 'in', 'what is', 'is the', 'to', 'what is the', 'was']

words: wordcount what : 1152, the : 876, of : 452, is : 400, in : 302, what is : 253, is the : 171, to : 163, what is the : 141, was : 138,

Category: ABBR words: ['what', 'for', 'the', 'does', 'what does', 'stand', 'stand for', 'is', 'what is', 'is the']

words: wordcount what : 86, for : 54, the : 48, does : 48, what does : 48, stand : 41, stand for : 41, is : 33, what is : 30, is the : 16,

Category: HUM words: ['the', 'who', 'what', 'of', 'in', 'was', 'is', 'of the', 'to', 'name']

words: wordcount the : 1019, who : 597, what : 547, of : 387, in : 299, was : 291, is : 234, of the : 156, to : 155, name : 144,

Category: NUM words: ['the', 'how', 'many', 'how many', 'what', 'of', 'in', 'is', 'when', 'is the']

words: wordcount the : 560, how : 492, many : 323, how many : 323, what : 276, of : 246, in : 245, is : 223, when : 136, is the : 136,

Category: LOC words: ['the', 'what', 'is', 'where', 'in', 'of', 'is the', 'country', 'city', 'what is']

words: wordcount the : 669, what : 554, is : 304, where : 258, in : 221, of : 201, is the : 139, country : 123, city : 101, what is : 95,

ngram_range: (2, 3)

Category: DESC words: ['what is', 'is the', 'what is the', 'how do', 'of the', 'what are', 'what does', 'do you', 'how do you', 'the origin']

words: wordcount what is : 420, is the : 194, what is the : 179, how do : 125, of the : 85, what are : 77, what does : 68, do you : 64, how do you : 58, the origin : 53,

Category: ENTY words: ['what is', 'is the', 'what is the', 'of the', 'what was', 'in the', 'what are', 'was the', 'what the', 'fear of']

words: wordcount what is : 253, is the : 171, what is the : 141, of the : 103, what was : 78, in the : 75, what are : 75, was the : 63, what the : 62, fear of : 62,

Category: ABBR words: ['what does', 'stand for', 'what is', 'is the', 'what is the', 'the abbreviation', 'does the', 'what does the', 'is the abbreviation', 'abbreviation for']

words: wordcount what does : 48, stand for : 41, what is : 30, is the : 16, what is the : 14, the abbreviation : 14, does the : 12, what does the : 12, is the abbreviation : 8, abbreviation for : 7,

Category: HUM words: ['of the', 'who was', 'was the', 'is the', 'who is', 'in the', 'who was the', 'what is', 'the first', 'who is the']

words: wordcount of the : 156, who was : 144, was the : 124, is the : 107, who is : 107, in the : 99, who was the : 84, what is : 62, the first : 61, who is the : 58,

Category: NUM words: ['how many', 'is the', 'what is', 'what is the', 'are there', 'was the', 'in the', 'how much', 'when was', 'how long']

words: wordcount how many : 323, is the : 136, what is : 111, what is the : 101, are there : 71, was the : 67, in the : 61, how much : 57, when was : 53, how long : 50,

Category: LOC words: ['is the', 'what is', 'what country', 'what is the', 'where is', 'in the', 'where can', 'of the', 'the world', 'what city']

words: wordcount is the : 139, what is : 95, what country : 86, what is the : 80, where is : 77, in the : 74, where can : 67, of the : 62, the world : 57, what city : 53,

ngram_range: (3, 3)

Category: DESC words: ['what is the', 'how do you', 'is the origin', 'the origin of', 'origin of the', 'what are the', 'the difference between', 'what does the', 'is the difference', 'what was the']

words: wordcount what is the : 179, how do you : 58, is the origin : 52, the origin of : 52, origin of the : 39, what are the : 35, the difference between : 30, what does the : 29, is the difference : 26, what was the : 20,

Category: ENTY words: ['what is the', 'what is fear', 'is fear of', 'the name of', 'what was the', 'what are the', 'name of the', 'what kind of', 'is the name', 'was the name']

words: wordcount what is the : 141, what is fear : 58, is fear of : 58, the name of : 57, what was the : 54, what are the : 47, name of the : 40, what kind of : 39, is the name : 28, was the name : 26,

Category: ABBR words: ['what is the', 'what does the', 'is the abbreviation', 'the abbreviation for', 'what does stand', 'does stand for', 'the national bureau', 'national bureau of', 'bureau of investigation', 'stand for in']

words: wordcount what is the : 14, what does the : 12, is the abbreviation : 8, the abbreviation for : 6, what does stand : 5, does stand for : 5, the national bureau : 5, national bureau of : 5, bureau of investigation : 5, stand for in : 5,

Category: HUM words: ['who was the', 'who is the', 'the name of', 'what is the', 'was the first', 'name of the', 'what was the', 'who wrote the', 'who invented the', 'is the name']

words: wordcount who was the : 84, who is the : 58, the name of : 42, what is the : 39, was the first : 35, name of the : 35, what was the : 32, who wrote the : 21, who invented the : 20, is the name : 19,

Category: NUM words: ['what is the', 'how many people', 'are there in', 'when was the', 'what year did', 'how long does', 'does it take', 'in what year', 'when did the', 'what year was']

words: wordcount what is the : 101, how many people : 35, are there in : 34, when was the : 30, what year did : 23, how long does : 20, does it take : 20, in what year : 20, when did the : 16, what year was : 15,

Category: LOC words: ['what is the', 'where can find', 'in the world', 'where is the', 'what are the', 'is the largest', 'what city is', 'what country is', 'the name of', 'where did the']

words: wordcount what is the : 80, where can find : 41, in the world : 34, where is the : 33, what are the : 23, is the largest : 21, what city is : 20, what country is : 15, the name of : 13, where did the : 13,

Text cleaned

ngram_range: (1, 1)

Category: DESC words: ['what', 'how', 'why', 'mean', 'origin', 'get', 'name', 'difference', 'word', 'find']

words: wordcount what : 749, how : 278, why : 104, mean : 62, origin : 54, get : 37, name : 32, difference : 32, word : 30, find : 27,

Category: ENTY words: ['what', 'name', 'fear', 'first', 'kind', 'which', 'called', 'used', 'world', 'film']

words: wordcount what : 1112, name : 119, fear : 66, first : 50, kind : 43, which : 43, called : 41, used : 38, world : 37, film : 36,

Category: ABBR words: ['what', 'stand', 'abbreviation', 'mean', 'national', 'bureau', 'investigation', 'acronym', 'used', 'cnn']

words: wordcount what : 81, stand : 41, abbreviation : 16, mean : 9, national : 5, bureau : 5, investigation : 5, acronym : 4, used : 3, cnn : 3,

Category: HUM words: ['who', 'what', 'name', 'the', 'first', 'president', 'which', 'company', 'wrote', 'world']

words: wordcount who : 559, what : 535, name : 144, the : 86, first : 82, president : 65, which : 46, company : 44, wrote : 38, world : 37,

Category: NUM words: ['how', 'many', 'what', 'when', 'year', 'much', 'long', 'people', 'first', 'take']

words: wordcount how : 479, many : 323, what : 245, when : 124, year : 68, much : 57, long : 56, people : 41, first : 35, take : 31,

Category: LOC words: ['what', 'where', 'country', 'city', 'state', 'world', 'find', 'largest', 'name', 'river']

words: wordcount what : 524, where : 255, country : 123, city : 102, state : 65, world : 62, find : 52, largest : 46, name : 32, river : 26,

ngram_range: (1, 2)

Category: DESC words: ['what', 'how', 'why', 'mean', 'origin', 'what origin', 'get', 'name', 'difference', 'what difference']

words: wordcount what : 749, how : 278, why : 104, mean : 62, origin : 54, what origin : 53, get : 37, name : 32, difference : 32, what difference : 32,

Category: ENTY words: ['what', 'name', 'fear', 'what fear', 'what name', 'first', 'kind', 'which', 'called', 'used']

words: wordcount what : 1112, name : 119, fear : 66, what fear : 62, what name : 60, first : 50, kind : 43, which : 43, called : 41, used : 38,

Category: ABBR words: ['what', 'stand', 'abbreviation', 'what abbreviation', 'mean', 'what stand', 'national', 'bureau', 'investigation', 'national bureau']

words: wordcount what : 81, stand : 41, abbreviation : 16, what abbreviation : 13, mean : 9, what stand : 8, national : 5, bureau : 5, investigation : 5, national bureau : 5,

Category: HUM words: ['who', 'what', 'name', 'the', 'first', 'president', 'which', 'company', 'what name', 'wrote']

words: wordcount who : 559, what : 535, name : 144, the : 86, first : 82, president : 65, which : 46, company : 44, what name : 42, wrote : 38,

Category: NUM words: ['how', 'many', 'how many', 'what', 'when', 'year', 'much', 'long', 'how much', 'how long']

words: wordcount how : 479, many : 323, how many : 316, what : 245, when : 124, year : 68, much : 57, long : 56, how much : 54, how long : 48,

Category: LOC words: ['what', 'where', 'country', 'city', 'what country', 'state', 'world', 'find', 'largest', 'where find']

words: wordcount what : 524, where : 255, country : 123, city : 102, what country : 86, state : 65, world : 62, find : 52, largest : 46, where find : 45,

ngram_range: (2, 2)

Category: DESC words: ['what origin', 'what difference', 'how find', 'how get', 'origin word', 'what causes', 'what meaning', 'what name', 'what definition', 'what history']

words: wordcount what origin : 53, what difference : 32, how find : 25, how get : 22, origin word : 17, what causes : 16, what meaning : 12, what name : 12, what definition : 11, what history : 9,

Category: ENTY words: ['what fear', 'what name', 'what kind', 'what color', 'what first', 'what film', 'what term', 'what common', 'what animal', 'soft drink']

words: wordcount what fear : 62, what name : 60, what kind : 37, what color : 27, what first : 15, what film : 14, what term : 13, what common : 13, what animal : 13, soft drink : 13,

Category: ABBR words: ['what abbreviation', 'what stand', 'national bureau', 'bureau investigation', 'what abbreviated', 'what letters', 'what acronym', 'what full', 'general motors', 'what ioc']

words: wordcount what abbreviation : 13, what stand : 8, national bureau : 5, bureau investigation : 5, what abbreviated : 3, what letters : 3, what acronym : 3, what full : 2, general motors : 2, what ioc : 2,

Category: HUM words: ['what name', 'who first', 'who wrote', 'who invented', 'what president', 'who portrayed', 'who played', 'what company', 'what famous', 'who created']

words: wordcount what name : 42, who first : 33, who wrote : 29, who invented : 28, what president : 18, who portrayed : 17, who played : 15, what company : 14, what famous : 14, who created : 13,

Category: NUM words: ['how many', 'how much', 'how long', 'many people', 'what year', 'in year', 'what average', 'how old', 'long take', 'many years']

words: wordcount how many : 316, how much : 54, how long : 48, many people : 35, what year : 23, in year : 19, what average : 16, how old : 15, long take : 14, many years : 12,

Category: LOC words: ['what country', 'where find', 'what city', 'what state', 'what largest', 'what two', 'where get', 'what name', 'what capital', 'united states']

words: wordcount what country : 86, where find : 45, what city : 45, what state : 40, what largest : 20, what two : 17, where get : 13, what name : 12, what capital : 12, united states : 11,

ngram_range: (1, 3)

Category: DESC words: ['what', 'how', 'why', 'mean', 'origin', 'what origin', 'get', 'name', 'difference', 'what difference']

words: wordcount what : 749, how : 278, why : 104, mean : 62, origin : 54, what origin : 53, get : 37, name : 32, difference : 32, what difference : 32,

Category: ENTY words: ['what', 'name', 'fear', 'what fear', 'what name', 'first', 'kind', 'which', 'called', 'used']

words: wordcount what : 1112, name : 119, fear : 66, what fear : 62, what name : 60, first : 50, kind : 43, which : 43, called : 41, used : 38,

Category: ABBR words: ['what', 'stand', 'abbreviation', 'what abbreviation', 'mean', 'what stand', 'national', 'bureau', 'investigation', 'national bureau']

words: wordcount what : 81, stand : 41, abbreviation : 16, what abbreviation : 13, mean : 9, what stand : 8, national : 5, bureau : 5, investigation : 5, national bureau : 5,

Category: HUM words: ['who', 'what', 'name', 'the', 'first', 'president', 'which', 'company', 'what name', 'wrote']

words: wordcount who : 559, what : 535, name : 144, the : 86, first : 82, president : 65, which : 46, company : 44, what name : 42, wrote : 38,

Category: NUM words: ['how', 'many', 'how many', 'what', 'when', 'year', 'much', 'long', 'how much', 'how long']

words: wordcount how : 479, many : 323, how many : 316, what : 245, when : 124, year : 68, much : 57, long : 56, how much : 54, how long : 48,

Category: LOC words: ['what', 'where', 'country', 'city', 'what country', 'state', 'world', 'find', 'largest', 'where find']

words: wordcount what : 524, where : 255, country : 123, city : 102, what country : 86, state : 65, world : 62, find : 52, largest : 46, where find : 45,

ngram_range: (2, 3)

Category: DESC words: ['what origin', 'what difference', 'how find', 'how get', 'origin word', 'what origin word', 'what causes', 'what meaning', 'what name', 'what definition']

words: wordcount what origin : 53, what difference : 32, how find : 25, how get : 22, origin word : 17, what origin word : 17, what causes : 16, what meaning : 12, what name : 12, what definition : 11,

Category: ENTY words: ['what fear', 'what name', 'what kind', 'what color', 'what first', 'what film', 'what term', 'what common', 'what animal', 'soft drink']

words: wordcount what fear : 62, what name : 60, what kind : 37, what color : 27, what first : 15, what film : 14, what term : 13, what common : 13, what animal : 13, soft drink : 13,

Category: ABBR words: ['what abbreviation', 'what stand', 'national bureau', 'bureau investigation', 'national bureau investigation', 'what abbreviated', 'what letters', 'what acronym', 'what full', 'general motors']

words: wordcount what abbreviation : 13, what stand : 8, national bureau : 5, bureau investigation : 5, national bureau investigation : 5, what abbreviated : 3, what letters : 3, what acronym : 3, what full : 2, general motors : 2,

Category: HUM words: ['what name', 'who first', 'who wrote', 'who invented', 'what president', 'who portrayed', 'who played', 'what company', 'what famous', 'who created']

words: wordcount what name : 42, who first : 33, who wrote : 29, who invented : 28, what president : 18, who portrayed : 17, who played : 15, what company : 14, what famous : 14, who created : 13,

Category: NUM words: ['how many', 'how much', 'how long', 'many people', 'how many people', 'what year', 'in year', 'what average', 'how old', 'long take']

words: wordcount how many : 316, how much : 54, how long : 48, many people : 35, how many people : 35, what year : 23, in year : 19, what average : 16, how old : 15, long take : 14,

Category: LOC words: ['what country', 'where find', 'what city', 'what state', 'what largest', 'what two', 'where get', 'what name', 'what capital', 'united states']

words: wordcount what country : 86, where find : 45, what city : 45, what state : 40, what largest : 20, what two : 17, where get : 13, what name : 12, what capital : 12, united states : 11,

ngram_range: (3, 3)

Category: DESC words: ['what origin word', 'what claim fame', 'what origin term', 'what origin name', 'san diego schools', 'what meaning name', 'jane goodall famous', 'how find list', 'why people get', 'what dew point']

words: wordcount what origin word : 17, what claim fame : 5, what origin term : 4, what origin name : 3, san diego schools : 3, what meaning name : 3, jane goodall famous : 3, how find list : 2, why people get : 2, what dew point : 2,

Category: ENTY words: ['what soft drink', 'world war ii', 'first singing cowboy', 'what best way', 'ibm compatible machines', 'what tv series', 'islamic counterpart red', 'counterpart red cross', 'stuart hamblen considered', 'hamblen considered first']

words: wordcount what soft drink : 7, world war ii : 6, first singing cowboy : 5, what best way : 4, ibm compatible machines : 4, what tv series : 4, islamic counterpart red : 4, counterpart red cross : 4, stuart hamblen considered : 3, hamblen considered first : 3,

Category: ABBR words: ['national bureau investigation', 'what stand equation', 'stand equation mc2', 'what full form', 'full form com', 'what abbreviation aids', 'abbreviation aids stand', 'what inri stand', 'inri stand used', 'stand used jesus']

words: wordcount national bureau investigation : 5, what stand equation : 2, stand equation mc2 : 2, what full form : 1, full form com : 1, what abbreviation aids : 1, abbreviation aids stand : 1, what inri stand : 1, inri stand used : 1, stand used jesus : 1,

Category: HUM words: ['who wrote the', 'randy steven craft', 'who first woman', 'lawyer represented randy', 'who made first', 'world war ii', 'who first american', 'who prime minister', 'what real name', 'who first black']

words: wordcount who wrote the : 10, randy steven craft : 6, who first woman : 5, lawyer represented randy : 5, who made first : 5, world war ii : 5, who first american : 4, who prime minister : 4, what real name : 4, who first black : 4,

Category: NUM words: ['how many people', 'how long take', 'how many years', 'how many times', 'how much money', 'how many miles', 'many people died', 'many people randy', 'how many different', 'watch the simpsons']

words: wordcount how many people : 35, how long take : 14, how many years : 11, how many times : 11, how much money : 7, how many miles : 7, many people died : 6, many people randy : 5, how many different : 5, watch the simpsons : 5,

Category: LOC words: ['where find information', 'what two countries', 'what largest city', 'kentucky horse park', 'what country boasts', 'what south american', 'what european country', 'where get information', 'what country capital', 'city maurizio pellegrin']

words: wordcount where find information : 9, what two countries : 8, what largest city : 7, kentucky horse park : 6, what country boasts : 5, what south american : 4, what european country : 4, where get information : 4, what country capital : 3, city maurizio pellegrin : 3,