Tamir Moshiashvili

Amos Mimon

Report:

1. Statistics:  
   a. threshold:  
    - filtered the words that appeared below 100 times (those words has no vector)  
   b. number of words that occurred at least 100 times - 14490  
   c. number of features for each word – at most, 50
2. 20-most similar words for each target-word is in a separate file, called "most\_common\_words.csv".  
   \* Conclusions for co-occurrence-1:  
    similar words has the same topic (topic related), that's because here, the features  
    for each word  
    are the words in the same sentence.  
    By that we can assume to have strong relation between the topic to each word.  
   \* Conclusions for co-occurrence-2:  
    similar words are the words with the same topic (topic related) and has the same  
    role in the sentence (semantic class) because of the features being picked from the  
    window.  
    In addition, ignoring functional-words strengthen the similarity between words  
    which are topic-related.  
   \* Conclusions for co-occurrence-3:  
    similar words here have the same role in the sentence (semantic related)  
    because the features been taken from dependency-edges.
3. For car:

Topic-related marked in **red**

Semantic-related marked in **blue**

TODO

|  |  |  |  |
| --- | --- | --- | --- |
| car |  |  |  |
|  | andretti | auto | automobile |
|  | bmw | automobile | bike |
|  | chevrolet | champ | boat |
|  | fia | constructor | carriage |
|  | ford | driver | chassis |
|  | gt | endurance | cruiser |
|  | harvick | fia | diesel |
|  | midget | formula | driver |
|  | motorcycle | indy | ferry |
|  | motorsport | lap | motor |
|  | motorsports | motor | motorcycle |
|  | nascar | motorcycle | operator |
|  | nissan | nascar | passenger |
|  | suv | off-road | race |
|  | tdi | racing | racing |
|  | toyota | stock | terminal |
|  | traction | thoroughbred | traffic |
|  | truck | toyota | truck |
|  | vauxhall | truck | vehicle |
|  | waltrip | yacht | wagon |

For piano:

Topic-related marked in **red**

Semantic-related marked in **blue**

TODO

|  |  |  |  |
| --- | --- | --- | --- |
| piano |  |  |  |
|  | banjo | beethoven | bass |
|  | bassoon | cello | concerto |
|  | beethoven | clarinet | drum |
|  | cello | concerto | duo |
|  | clarinet | flute | flute |
|  | concerto | guitar | guitar |
|  | continuo | handel | guitarist |
|  | flute | harp | instrument |
|  | harp | mozart | keyboard |
|  | harpsichord | oboe | lesson |
|  | mandolin | orchestra | op |
|  | oboe | recital | playing |
|  | op | saxophone | quartet |
|  | quintet | soloist | solo |
|  | sonata | sonata | soloist |
|  | timpani | soprano | sonata |
|  | trombone | trombone | trio |
|  | viola | trumpet | trumpet |
|  | violin | viola | violin |
|  | violoncello | violin | violinist |

Word2vec:

1. The 20-most-common-words for each target word is in a separate file called  
   "word2vec\_most\_common.csv"