Tamir Moshiashvili

Amos Maimon

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 are at most semantic-related because when the features are the  
    sentence's words, there might be a lot of features that can be used like the word,  
    thus the similar words has the same semantics.  
    But, also the features will contain topic-related words, so sometimes there will be  
    words that connected by topic.  
     
    For example, for car, most of the similar words are types of cars or a vehicle,  
    and for piano, there are few words that connected topically.  
     
   \* Conclusions for co-occurrence-2:  
    similar words may be both semantic and topic related because the features taken  
    from the window, so topic-related words are in the window (because we  
    ignore functional-words), but also semantic-related words (words that are close  
    to the target word in the sentence).  
     
    For example, in both car and piano, the amount of topic-related and  
    semantic-related is almost the same.  
     
   \* Conclusions for co-occurrence-3:  
    similar words are at most topic-related words, because the features are taken  
    from the dependency-tree, there is more chance that the features will be  
    connected by that topically.  
    But there might happen that a word will contain as feature a word that has the  
    semantics.  
     
    For example, in car, most of the similar words are topic-related, and there are  
    few words that are semantic-related,  
    and for piano, there is same amount for both.
3. 20-most common features for each target word in each co-occurrence is in a separate file, named "most\_common\_features.csv"  
   TODO write short comparisons between the table of features and table of similar words of section-2
4. For car:

Topic-related marked in **red**

Semantic-related marked in **blue**

|  |  |  |  |
| --- | --- | --- | --- |
| car | Co-occur1 | Co-occur2 | Co-occur3 |
|  | andretti | auto | airplane |
|  | bmw | automobile | amusement |
|  | chevrolet | champ | auto |
|  | fia | constructor | automobile |
|  | ford | driver | challenger |
|  | gt | endurance | deadly |
|  | harvick | fia | drunk |
|  | midget | formula | fatal |
|  | motorcycle | indy | freak |
|  | motorsport | lap | hunting |
|  | motorsports | motor | indoor |
|  | nascar | motorcycle | motorcycle |
|  | nissan | nascar | plane |
|  | suv | off-road | racing |
|  | tdi | racing | rally |
|  | toyota | stock | skiing |
|  | traction | thoroughbred | traffic |
|  | truck | toyota | tragic |
|  | vauxhall | truck | truck |
|  | waltrip | yacht | turbo |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Co-occurrence-1 | | Co-occurrence-2 | | Co-occurrence-3 | |
| topic | semantic | Topic | Semantic | topic | semantic |
| 2/20 | 18/20 | 7/20 | 12/20 | 13/20 | 7/20 |
| 10% | 90% | 35% | 60% | 65% | 35% |

Insights:

Most semantic-related words are type of a car or transportation-object.

Topic-related words are words that connected to car in different ways, like traffic, lap, drunk and so on.

For piano:

Topic-related marked in **red**

Semantic-related marked in **blue**

|  |  |  |  |
| --- | --- | --- | --- |
| piano | Co-occur1 | Co-occur2 | Co-occur3 |
|  | banjo | beethoven | banjo |
|  | bassoon | cello | bass |
|  | beethoven | clarinet | beethoven |
|  | cello | concerto | cello |
|  | clarinet | flute | clarinet |
|  | concerto | guitar | classical |
|  | continuo | handel | drum |
|  | flute | harp | flute |
|  | harp | mozart | guitar |
|  | harpsichord | oboe | harp |
|  | mandolin | orchestra | keyboard |
|  | oboe | recital | melodic |
|  | op | saxophone | mozart |
|  | quintet | soloist | professionally |
|  | sonata | sonata | sang |
|  | timpani | soprano | saxophone |
|  | trombone | trombone | strauss |
|  | viola | trumpet | tenor |
|  | violin | viola | viola |
|  | violoncello | violin | violin |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Co-occurrence-1 | | Co-occurrence-2 | | Co-occurrence-3 | |
| topic | semantic | Topic | Semantic | topic | semantic |
| 7/20% | 13/20 | 8/20 | 12/20 | 10/20 | 10/20 |
| 35% | 65% | 40% | 60% | 50% | 50% |

Insights:

Most semantic-related words are musical-instruments.

Topic related words are words from the music-world, like Beethoven and Mozart as composers, and sonata and concerto.

1. TODO write brief description on implementation

Word2vec:

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