**Semantic Analysis**

*Assignment 3*

*Computational Linguistics*

Cristina Luengo Agulló

Javier Navarro González

École Polytechnique de Louvain

Université Catholique de Louvain

December 09, 2014

**Introduction**

**Lexicon**

After removing the 250 most frequent words from the 5-gram corpus, we obtain the **lexicon** that we are going to use during the assignment. These are the top 20 most frequent words:

|  |  |
| --- | --- |
| Words | Frequency |
| home | 33.482 |
| reason | 33.138 |
| news | 32.772 |
| second | 32.645 |
| difficult | 32.014 |
| job | 31.782 |
| until | 31.475 |
| own | 30.931 |
| pay | 30.586 |
| war | 30.554 |
| big | 30.459 |
| looking | 30.190 |
| white | 30.155 |
| hands | 29.849 |
| show | 29.745 |
| women | 29.717 |
| order | 29.627 |
| beginning | 29.578 |
| different | 29.180 |
| ago | 29.165 |

Another method we could have used to discard the 250 most frequent words is by using the *idf* computation of each word. The *‘inverse document frequency’* is a term that measures how common a word is among a collection of documents, so the more frequent a word is, the lower is its *idf* value. Therefore, one way to determine which the 250 most frequent words are would be to calculate the idf value for each word of the lexicon and then sorting the words by their value in increasing order. The first 250 words will be the most frequent ones, so deleting them will give the requested lexicon.

**Bag of words**

Fireworks:

july: 10

fourth: 21

Furnace:

water: 6

heater: 6

**TF-IDF Representation**

Fireworks:

july: 26.5342

fourth: 46.9439

Furnace

water: 10.5198

heater: 24.5565

**Bonus task**

**Distribution of similarity scores**