Text Categorization project

Oussama Ennafii, Sammy Khalife & Shuyu Dong

ENS Cachan - Master MVA

April 22, 2015

Introduction:

- Reuters-21578 R8 data, 5485 samples train set, 2189 samples test set
- 8 class of documents to class using classical (bag of words) and new techniques (Graph of Words)

Structure of code:

The classification is done in three steps:

(1) Constructing the document term matrix: to get this matrix we need three functions: the first one is to construct a graph of words from a document. The second gets a document word matrix in which each word in the corpus has a weight in each document. The third fucntion constructs the document term matrix corresponding to the TW-IDF measure.

(2) Reducing the dimension of the document term matrix using either LSI or Chi-Square.

(3) Learning over the train data set using the SVM or AdaBoost.

Experiments:

- We did tried only a size window of 4.
- The unweighted directed graph yields better results than the other possible graphs. The difference is not that big though.
- As suspected SVM is well suited for high dimensionnal problems so it does perform better.
- We project the data matrix into a 100 dimension space using LSI. It does yield the highest resutls.
- We had two choices either learn on the train set and test on the other set separatly, the problem would be that there are words that may not be common. As we will see there is a difference.

Results:

- separatly:
  - Microaveraging
    - precision: 0.640475102787
    - recall: 0.640475102787
  - Macroaveraging
    - precision: 0.170622860994
    - recall: 0.203406087815
- jointly:
  - Microaveraging
    - precision: 0.667428049338
    - recall: 0.667428049338
  - Macroaveraging
    - precision: 0.631236989331
    - recall: 0.229378369

Bag of words - Method

- Preprocessing on text data to build Document-term matrix
- Latent semantic indexing, which performs SVD to the Document-term matrix
- Training of a classifier (5185 samples)
- Get score on the test data (2178 samples)

Bag of words - Experiments

- LSI projection space dimension : 100 ($\text{Score}_{100} > max(\text{score}_{50}, \text{score}_{200})$)
- Support Vector Machines kept (following Graph-of-word and TW-IDF: new approach to ad hoc IR)

Results

- "Jointly" LSI
  - Micro-averaging precision : 0.781635449977
  - Macro-averaging precision : 0.198438628628
  - Micro-averaging recall : 0.781635449977
  - Macro-averaging recall : 0.242023076066

Conclusion

- Bag of words representation does better in general in comparison with the graph of words representation except for the macro-averaging precision
- We should investigate in the future, the effect of a global property graph like PageRank