3D Semantic knowldege retrieval

Eduardo San Miguel Martín

I. Introduction

- LSA is a machine-learning model:
  - that induces representations of the meaning of words
  - by analyzing the relation between words and passages in large bodies of text (Corpus)

I. Introduction

- The method used to capture the essential semantic information is dimension reduction:
  - selecting the most important dimensions from a co-occurrence matrix decomposed using Singular Value Decomposition. (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1991)

I. Introduction

- As a result, LSA offers a way of assessing semantic similarity between any two samples of text in an automatic, unsupervised way. (Landauer & Dumais, 1997).

I. Introduction

- LSA has been used in applied settings with a high degree of success in areas like:
  - automatic essay grading

  (Foltz, Laham, & Landauer, 1999)
  - automatic tutoring to improve summarization skills in children

  (E. Kintsch, Steinhart, Stahl, Matthews, Lamb, & the LSA Research Group, 2000).

I. Introduction

- As a model LSA's most impressive achievements have been:
  - in human language acquisition simulations (Landauer & Dumais, 1997)
  - and modeling of high-level comprehension phenomena like metaphor understanding, causal inferences and judgments of similarity (Kintsch, 2001).

I. Introduction

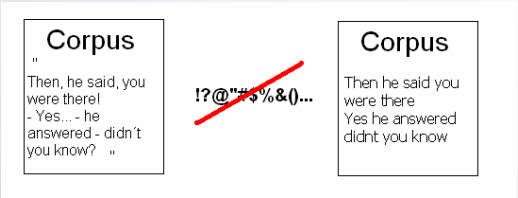
- Due to multidimensional nature of semantic space, LSA results are hard to visualize.
- Our goal was to develop a tool for visualizing this semantic relationship.

II. LSA implementation in R

Algorithm developed in R:

1. scan text:
   - Corpus with 6000 diferent words over 1000 documents
   - From "The catcher in the rye"

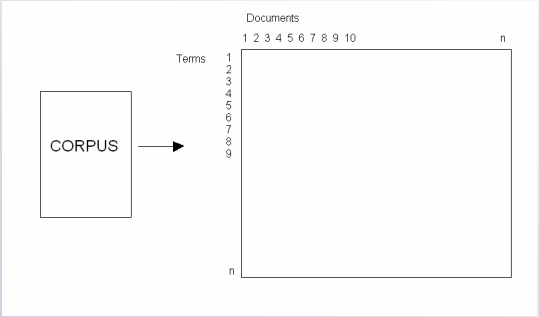
II. LSA implementation in R



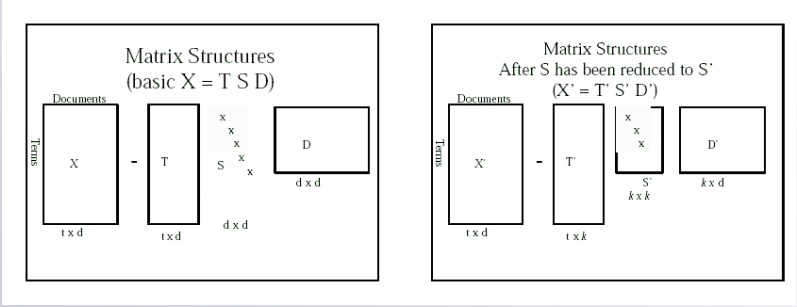
- 2. normalize text:
  - remove puntuation
  - word stemming using Porter´s algortihm

II. LSA implementation in R

- 3. Build Word by Document matrix



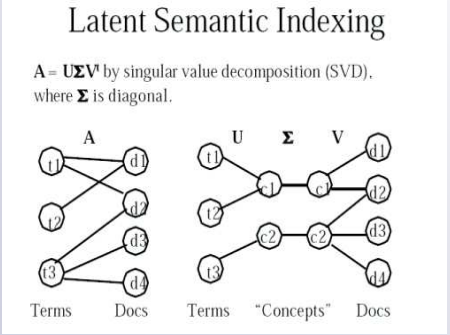
II. LSA implementation in R



- 4. Singular Value Decomposition

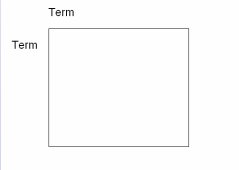
II. LSA implementation in R

- 5. Build new low dimensional matrix



II. LSA implementation in R

- 6.a Build term by term distances matrices:

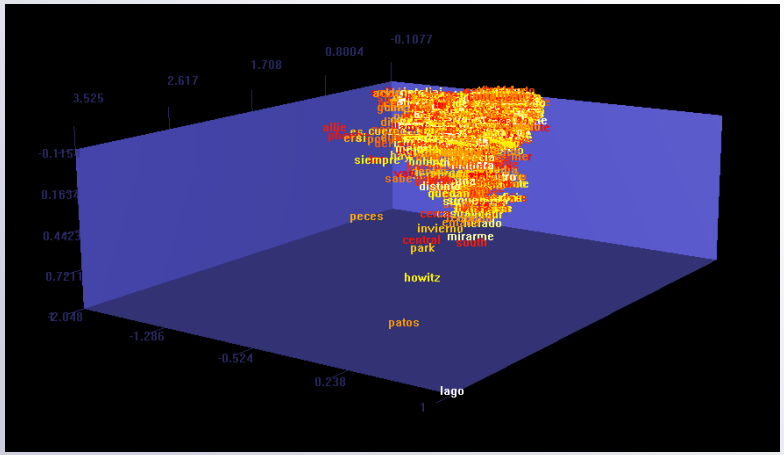


– Cosin, length & 1/eucl

III. Results

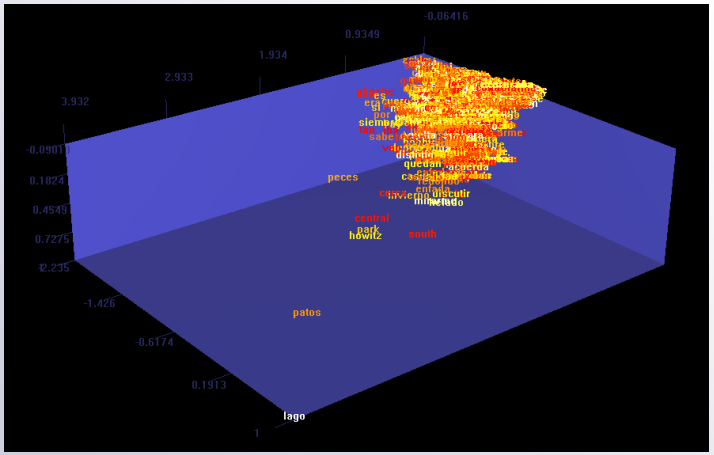
- 3D plots for word “Lago” (Lake)
  - Dimension reduction:
    - 100
    - 200

III. Results



- 3d plot for word Lake in a 100 dimensional space

III. Results

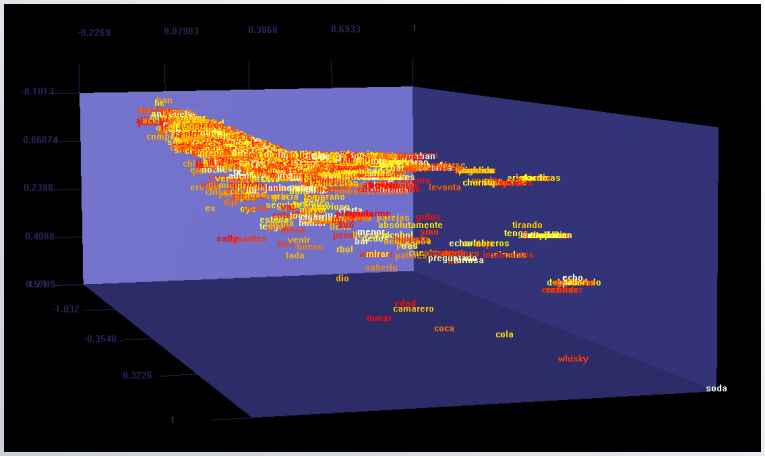


- 3d plot for word Lake in a 200 dimensional space

III. Results

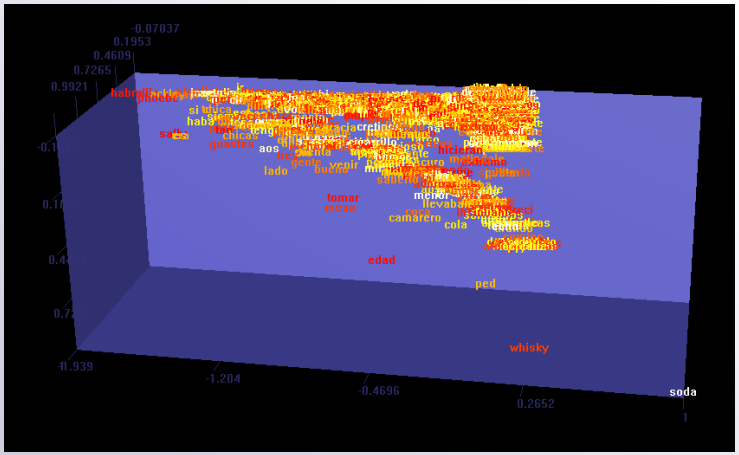
- 3D plots for word “Soda”
  - Dimension reduction:
    - 100
    - 200

III. Results



3d plot for word “Soda” in a 100 dimensional space

III. Results



3d plot for word “Soda” in a 200 dimensional space

IV. Conclusion

- R offers an enviroment for lsa visualization.
- Applications:
  - Psychology
  - Linguistics
  - Cognitive Science
  - User modelling
  - Etc…
- Future work:
  - Dimension reduction using cmdscale has offer ad-hoc results in some cases.

Literature cited

- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1991). Indexing By Latent Semantic Analysis. *Journal of the American Society For Information Science, 41*,391-407.
- Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Education Journal of Computer enhanced learning On-line journal, 1(2)*.
- Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., Lamb, R., & the LSA research Group (2000). Developing summarization skills through the use of LSA-backed feedback. *Interactive Learning Environments, 8*(2), 87-109.
- Kintsch, W. (2001). Predication. *Cognitive Science, 25*, 173-202.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.