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)

• 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)

• 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)

II. LSA implementation in R

Algorithm developed in R:
1. scan text:
   – Corpus with 6000 different words over 1000 documents
   – From “The catcher in the rye”

• 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).

• 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

• 2. normalize text:
  – remove punctuation
  – word stemming using Porter’s algorithm

• 3. Build Word by Document matrix

• 4. Singular Value Decomposition

• 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

- 3d plot for word Lake in a 100 dimensional space
- 3d plot for word Lake in a 200 dimensional space
III. Results

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

IV. Conclusion

• R offers an environment for LSA visualization.

• Applications:
  – Psychology
  – Linguistics
  – Cognitive Science
  – User modeling
  – Etc...

• Future work:
  – Dimension reduction using cmdscale has offered ad-hoc results in some cases.
Literature cited