SVD, LSA and LSI

Dr. Goutam Chakraborty

Objectives

- Explain the conceptual and mathematical basics of SVD, LSA and LSI in analyzing textual data.
- Explain the terms and document mapping in the SVD space.

What is SVD?

- **SVD as a Factorization Technique**
  - Singular value decomposition (SVD) is a means of decomposing a matrix into *a product of three simpler* matrices. This is how it is typically used in text analytics
  - In this way it is related to other matrix decompositions such as eigen decomposition, principal components analysis (PCA), and non-negative matrix factorization (NNMF).
  - Typically, term-by-document matrix is very large and sparse and rarely used directly in documents clustering or for topic extraction from documents.
    - Instead, the matrix is transformed to reduce its dimensionality (via SVD) yet retain most of the information.
    - Then SVD numbers are used for clustering, topic mining etc.
- SVD may also be used in other applications such as “least squares approximation” or, “for summarizing matrices with partial values”

SVD, LSA and LSI

- **Latent Semantic Analysis:** LSA is an application of **reduced-order SVD** in which the rows of the input matrix represent words and the columns documents, with entries being the count of the words in the document.
  - The singular vectors and corresponding singular values produced by SVD allow words and documents to be mapped into the same "latent semantic space".
  - The resulting embedding places similar words and documents as measured by co-occurrence near one another even if they never co-occurred in the training corpus.
- **Latent Semantic Indexing :** Application of LSA's notion of term-document similarity to information retrieval, the resulting systems being known as latent semantic indexing (LSI).
  - A query consisting of several terms corresponds to the sum of their k-dimensional vectors. The resulting k-dimensional query vector may be compared to the k-dimensional document vectors to determine similarity (typically using the *cosine similarity measure*).

Mathematics of SVD

- Let **A** (mxn) be the term-by-document matrix with m>n (*more terms* than documents) and where the entries in the matrix are real numbers (such as presence or absence of a term, or entropy weights).
- SVD will compute matrices **U**, **S**, and **V** such that the original matrix can be re-created using the formula,
- $\mathbf{A} = \mathbf{USV}^T$, where:
  - U is the matrix of orthogonal Eigenvectors of the square symmetric matrix $\mathbf{AA}^T$.
  - S is the diagonal matrix of square roots of nonzero eigenvalues of the square symmetric matrix $\mathbf{AA}^T$.
  - V is the matrix of orthogonal eigenvectors of the square symmetric matrix $\mathbf{A}^T\mathbf{A}$.

A Numerical Example of SVD

- Consider these **4** text documents.
- D1: I love Ipad
- D2: Ipad is great for kids
- D3: Kids love to play soccer
- D4: I play soccer at OSU
- The stop words are underlined.

Term/Document	D1	D2	D3	D4
I	1	0	0	1
Love	1	0	1	0
Ipad	1	1	0	0
Is	0	1	0	1
Great	0	1	0	0
Kids	0	1	1	0
Play	0	0	1	1
Soccer	0	0	1	1
OSU	0	0	0	1

- **A is a (9X4)** term-by-document matrix
- **A^T is a (4X9)** document-by-term matrix
- **$A^T A$ is a (4X4)** document-by-document matrix
- **AA^T is a (9X9)** term-by-term matrix

Eigenvalues and Eigenvectors of $\mathbf{A}^T\mathbf{A}$

- The eigenvalues of $\mathbf{A}^T\mathbf{A}$ are 7.783, 3.511, 2.253 and 2.453.
- The square roots of the eigenvalues of $\mathbf{A}^T\mathbf{A}$ are 2.790, 1.874, 1.566, and 1.501.

U:

0.355	0.120	-0.088	0.649
0.314	-0.056	0.677	0.205
0.265	-0.575	0.045	0.355
0.380	-0.164	-0.560	-0.069
0.145	-0.430	-0.214	-0.182
0.340	-0.340	0.205	-0.513
0.429	0.356	0.072	-0.220
0.429	0.356	0.072	-0.220
0.235	0.266	-0.346	0.112

S:

2.790	0.000	0.000	0.000
0.000	1.874	0.000	0.000
0.000	0.000	1.566	0.000
0.000	0.000	0.000	1.501

$\mathbf{V}^T$

0.335	0.405	0.542	0.655
-0.273	-0.806	0.168	0.498
0.405	-0.335	0.655	-0.542
0.806	-0.273	-0.498	0.168

Compare A versus USV^T

A

Term/ Document	D1	D2	D3	D4
I	1	0	0	1
Love	1	0	1	0
Ipad	1	1	0	0
Is	0	1	0	1
Great	0	1	0	0
Kids	0	1	1	0
Play	0	0	1	1
Soccer	0	0	1	1
OSU	0	0	0	1

USV^T

0.999	0.000	-0.001	0.999
1.000	0.000	0.998	-0.001
1.000	0.999	0.000	-0.001
0.000	0.999	0.000	0.999
0.000	1.000	0.001	0.000
0.001	1.000	1.001	0.001
-0.001	-0.001	0.999	1.000
-0.001	-0.001	0.999	1.000
0.000	0.000	0.001	1.000

More on SVD

- In full SVD, there is no dimensionality reduction and hence no loss of information.
- In practice, we use only the first few eigenvalues and keep the SVDs corresponding to those eigenvalues.
  - This means some loss of information but a gain of simple structure to represent data.
  - The documents are represented in the SVD space by a column vector of the matrix $\mathbf{V}^T$.
  - The terms are represented in the SVD space by the row vectors of the multiplication of the matrix **U** and the matrix **S**.

Two-Dimensional Representation

ID	Type	SVD1	SVD2
D1	Document	0.335	-0.273
D2	Document	0.405	-0.806
D3	Document	0.542	0.168
D4	Document	0.655	0.498
I	Term	0.99	0.225
Love	Term	0.876	-0.105
Ipad	Term	0.739	-1.078
is	Term	1.06	-0.307
Great	Term	0.405	-0.806
Kids	Term	0.949	-0.637
Play	Term	1.197	0.667
Soccer	Term	1.197	0.667
OSU	Term	0.656	0.498

Documents: V^T

0.335 0.405 0.542 0.655
-0.273 -0.806 0.168 0.498
~~0.405 -0.335 0.655 -0.542~~
~~0.806 -0.273 -0.498 0.168~~

Terms: U.S

U:

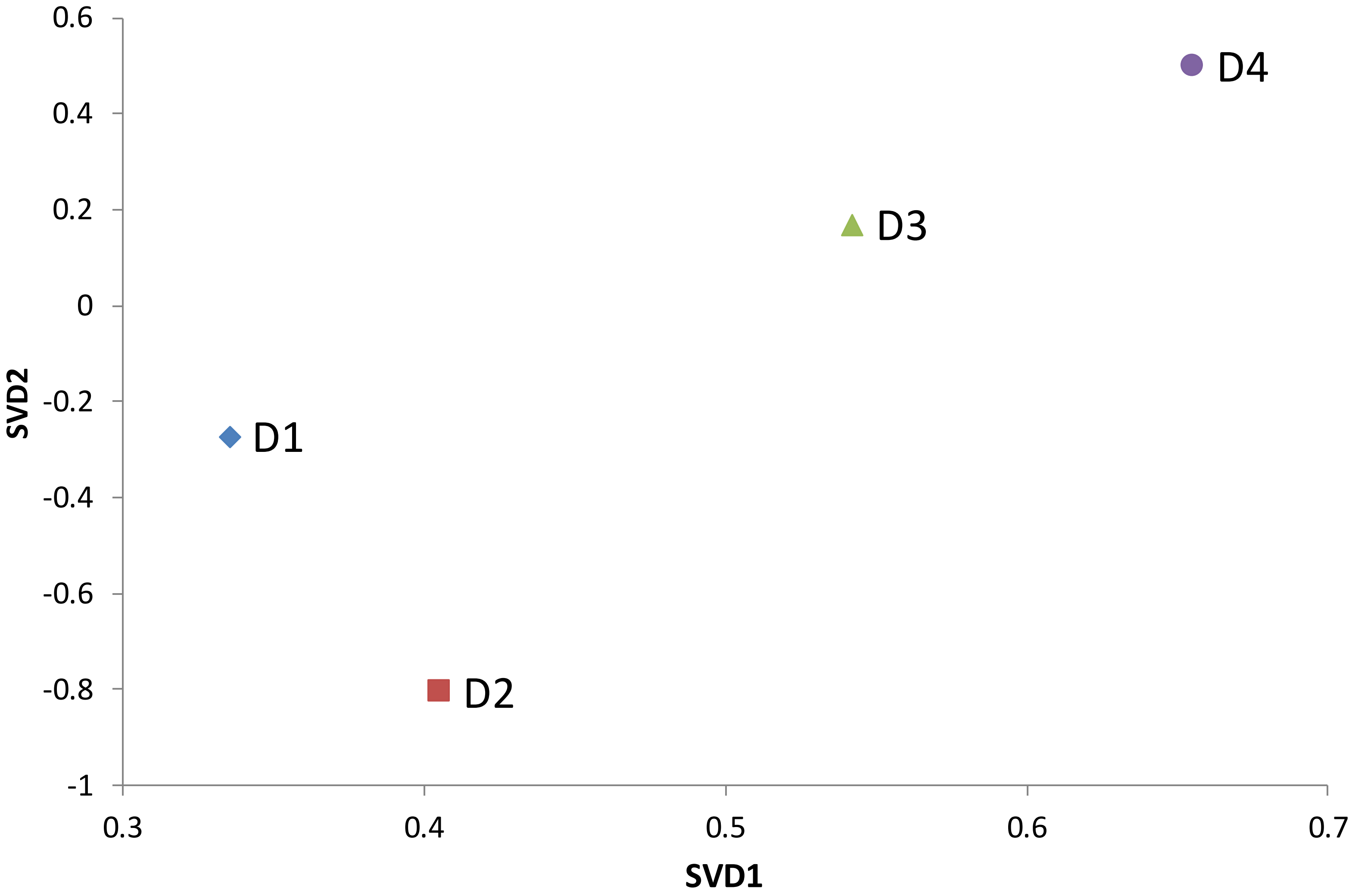
0.355 0.120 -0.088 0.649
0.314 -0.056 0.677 0.205
0.265 -0.575 0.045 0.355
0.380 -0.164 -0.560 -0.069
0.145 -0.430 -0.214 -0.182
0.340 -0.340 0.205 -0.513
0.429 0.356 0.072 -0.220
0.429 0.356 0.072 -0.220
0.235 0.266 -0.346 0.112

S:

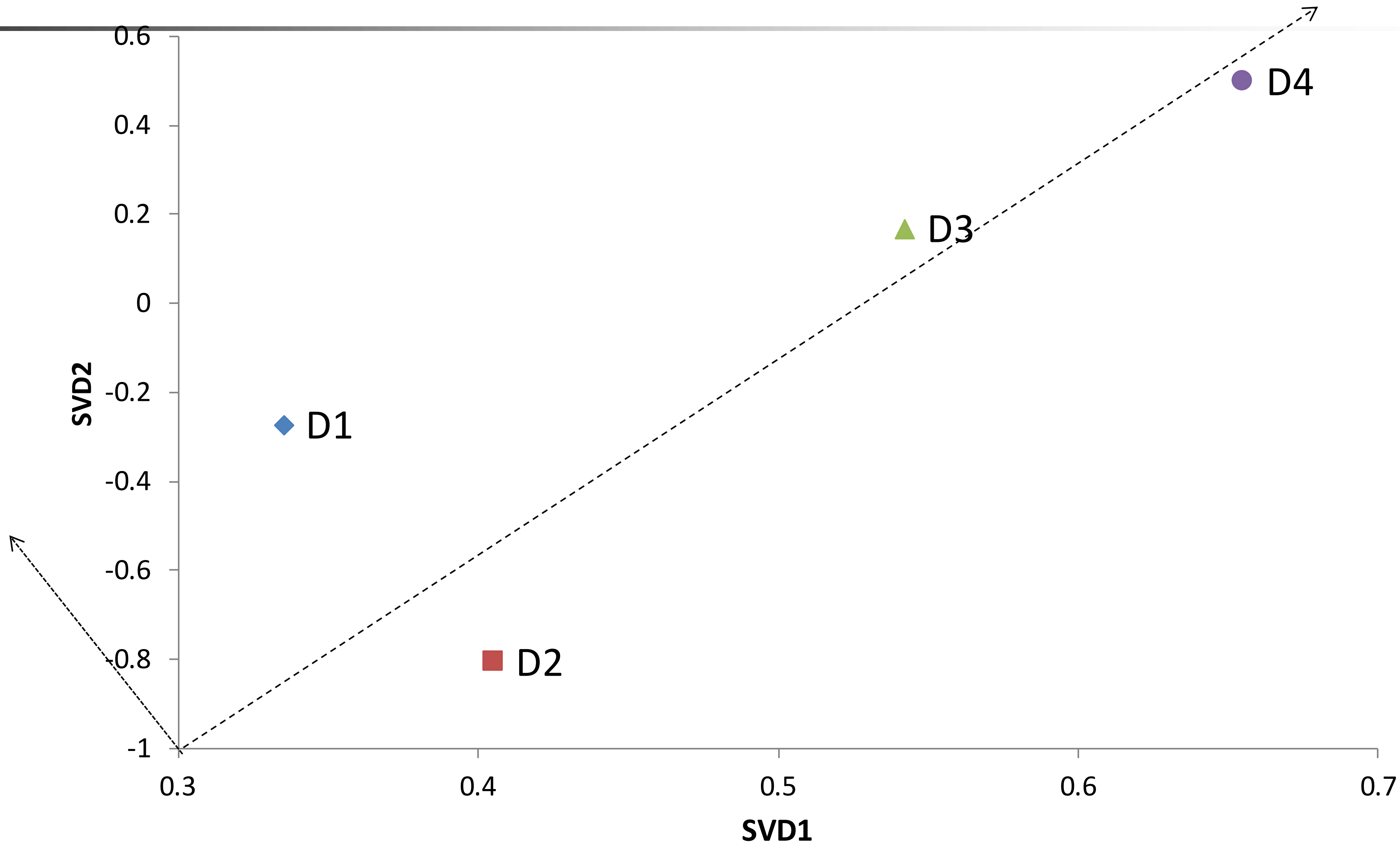
2.790 0.000 0.000 0.000
0.000 1.874 0.000 0.000
0.000 0.000 1.566 0.000
0.000 0.000 0.000 1.501

Term, I = 0.355*2.79 + 0.12*0 - 0.088*0 + 0.649*0= 0.99 (SVD 1)
Term, I = 0.324*0 - 0.056*1.874 + 0.677*0 + 0.205*0= -0.105 (SVD 2)

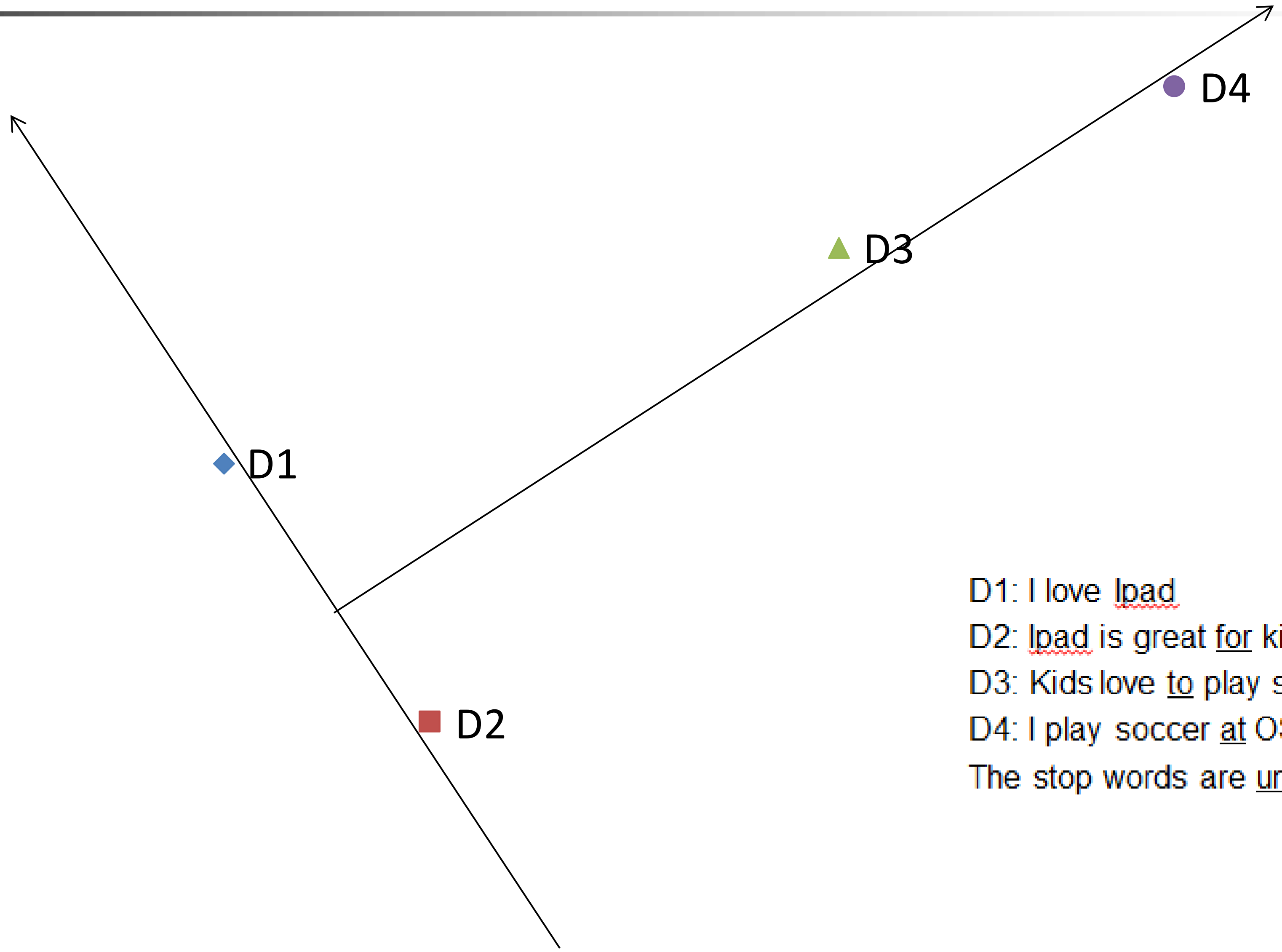
Document Plot in the SVD Space



Document Plot in the Rotated SVD Space



Document Plot in the Rotated SVD Space



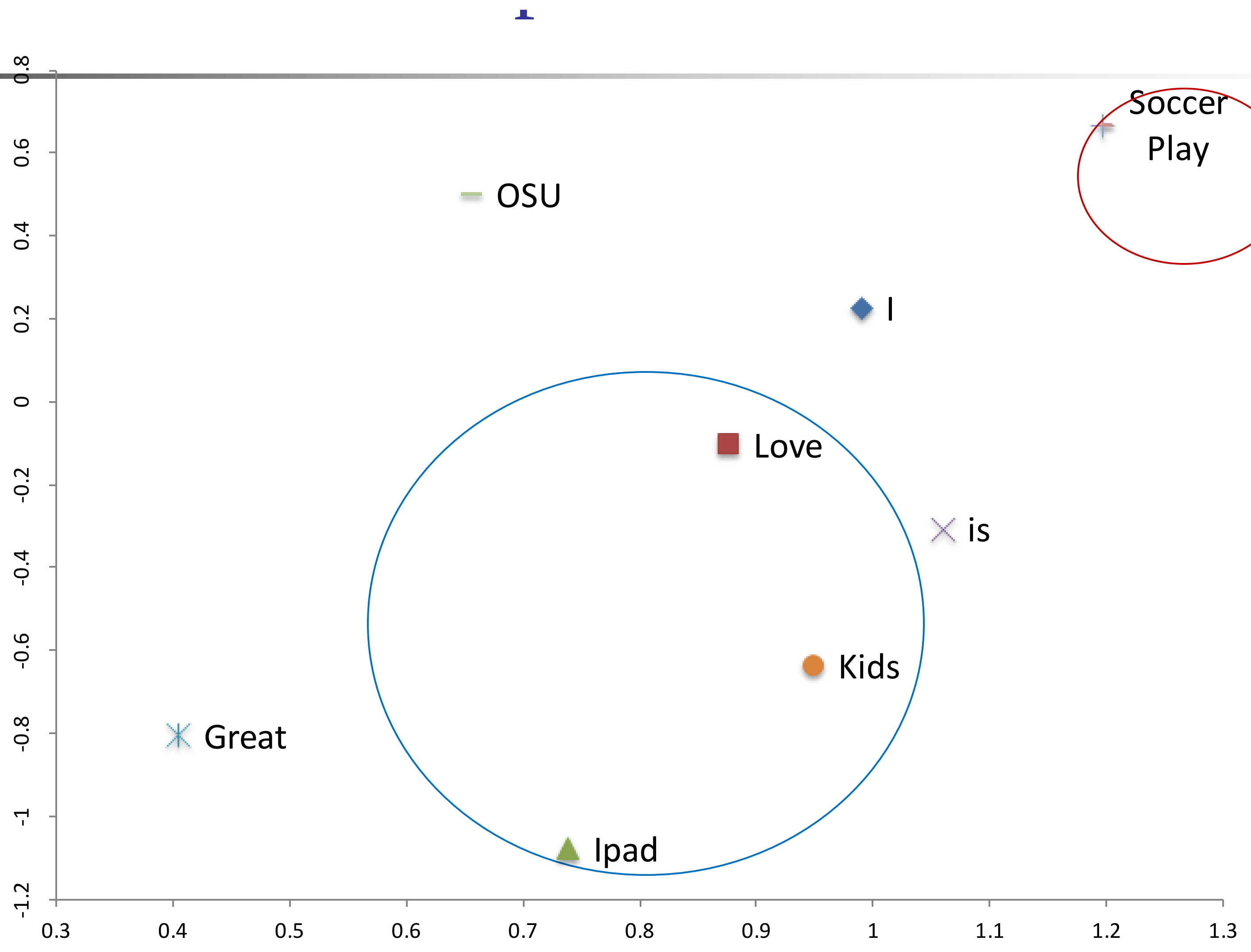
D1: I love Ipad

D2: Ipad is great for kids

D3: Kids love to play soccer

D4: I play soccer at OSU

The stop words are underlined.



Soccer
Play
OSU
I
Love
is
Kids
Great
Ipad
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3