Semantic Density Analysis: Comparing word meaning across time and phonetic space

Sagi, Kauffman, and Clark, Northwestern University

Paper Presentation
Text Mining: UVA Spring 2016
Hope McIntyre, Brian Sachtjen, Nick Venuti

Research Goal

It was a beautiful day in the neighborhood. The dog ran toward the fence.

I was walking the dog in the neighborhood. It started raining.

My friend passed by me. I said, “What up, dog?” He replied, “Not much.”

	...	dog	...
Doc1		1	
Doc2		1	
Doc3		1	

Challenges in Understanding Word Usage

- Word meanings have the tendency to vary
  - Multiple definitions
  - Different cultural norms
  - Temporal shifts
- Limited approaches to quantifying context
  - Lack of ordering in bag of words approach
  - Typically produce document level metrics (e.g. topical analysis)
  - Assumes word independence
  - Gives equal value for all occurrences of a word
  - Some words not present in manually annotated Lexicon

General Hypothesis for Quantifying Meaning

- The definition of a word can be gleaned from the words around it
- Word meanings can be compared by measuring the similarity of a word's contexts
- A greater context similarity = a smaller range in that word's meanings
- Compute **context vectors** to measure context similarity

Sagi, Kauffman, and Clark's Proposed Solution

1) **Word Vectors:** Develop co-occurrence matrix & reduce through Singular Value Decomposition
2) **Context Vectors:** Create context vectors based on value from co-occurrence matrix and words within k sized window
3) **Semantic Density:** Calculate average cosine similarities of context vectors

For Example:
Target Word:
"dog"
Target Window: 4

It was a beautiful day in the neighborhood. The dog ran toward the fence

I was walking the dog in the neighborhood. It started raining.

My friend passed by me. I said, "What up, dog?" He replied, "Not much."

Produce Word Vectors

It was a beautiful day in the neighborhood. The dog ran toward the fence.
I was walking the dog in the neighborhood. It started raining.
My friend passed by me. I said, “What up, dog?” He replied, “Not much.”

	a	beautiful	by	day	dog	fence	.	.	.
a	0	1	0	1	0	0	.	.	.
beautiful	1	0	0	1	0	0	.	.	.
by	0	0	0	0	0	0	.	.	.
day	1	1	0	0	0	0	.	.	.
dog	0	0	0	0	0	1	.	.	.
fence	0	0	0	0	1	0	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.

Produce Context Vectors

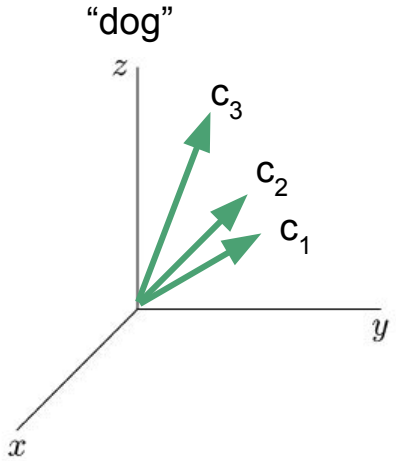
It was a beautiful day in the neighborhood. The dog ran toward the fence.
I was walking the dog in the neighborhood. It started raining.
My friend passed by me. I said, “What up, dog?” He replied, “Not much.”

	-4	-3	-2	-1	0	1	2	3	4
Sentence 1	in	the	neighborhood	the	dog	ran	toward	the	fence
Sentence 2	I	was	walking	the	dog	in	the	neighborhood	it
Sentence 3	I	said	what	up	dog	he	replied	not	much

	it	was	a	beautiful	day	in	the	neighborhood	dog	ran	toward	fence
it	..	..	..	..	..	..	..	..	..	..	..	..
was	..	..	..	..	..	..	..	..	..	..	..	..
a	..	..	..	..	..	..	..	..	..	..	..	..
beautiful	..	..	..	..	..	..	..	..	..	..	..	..
day	..	..	..	..	..	..	..	..	..	..	..	..
in	..	..	..	..	..	..	..	..	..	..	..	..
the	..	..	..	..	..	..	..	..	..	..	..	..
neighborhood	1	0	0	1	1	2	3	0	2	1	1	0
dog	..	..	..	..	..	..	..	..	..	..	..	..
ran	..	..	..	..	..	..	..	..	..	..	..	..
toward	..	..	..	..	..	..	..	..	..	..	..	..
fence	..	..	..	..	..	..	..	..	..	..	..	..

Calculate Target Word Semantic Density

- Density = Semantic variation within the set of individual occurrences of a given word, a more cohesive term has a higher density (word usage is “packed” in hyper-space)
- Measured by average cosine similarity $cos(\vec{w}, \vec{v}) = \frac{\vec{w} \cdot \vec{v}}{|\vec{w}||\vec{v}|}$



$$average\ cosine\ similarity = \frac{\cos(\vec{C}_1, \vec{C}_2) + \cos(\vec{C}_1, \vec{C}_3) + \cos(\vec{C}_2, \vec{C}_3)}{3}$$

Empirical Analysis

- Sagi et al. tested context vector methodology on Helsinki Corpus by investigating semantic shifts known from linguistic research
- Analyzed cases of semantic broadening, narrowing, and degeneration
- Ex. “Do”
  - Old English, used solely as a verb with a causative and habitual sense (e.g. “do you no harm”)
  - Later English, functional role, nearly devoid of meaning (e.g. “Do you know him?”)

	n	*Unknown composition date (<1250)*	*Early Middle English (1150-1350)*	*Late Middle English (1350-1500)*	*Early Modern English (1500-1710)*
dog	112			15.47 (14.19)	24.73(10.43)
do	4298		10.31(13.57)	13.02 (9.50)	24.54 (11.2)
deer	61	38.72 (17.59)	20.6 (18.18)		20.5 (9.82)
science	79			13.56 (13.33)	28.31 (12.24)

Limitations & Further Applications

- Target words need to be known or defined by experts
- High computational complexity
- Only useful for relative comparisons
- Still haven’t resolved all of the ambiguity of natural language
  - Word meaning depends on more than simple patterns of co-occurrence
- Further Applications:
  - Assist linguists in identifying new shifts in language trends
  - Predicting tendencies towards peace or violence in religious groups
  - Identify differences in word usage in American Presidential addresses
  - Cluster with these measurements to distinguish homonyms

Questions?