Lexicon, Syntax, Semantics IIb: Modeling Meaning

Distributional Semantics I

Eva Maria Vecchi

Center for Information and Language Processing
LMU Munich

May 28, 2020

The knowledge bottleneck

- Inference requires formalized *knowledge* about the world and about the meanings of words.
- **Q**: *Which genetically caused connective tissue disorder has severe symptoms and complications regarding the aorta and skeletal features, and, very characteristically, ophthalmologic subluxation?*
- **D**: *Marfan's is created by a defect of the gene that determines the structure of Fibrillin-11. One of the symptoms is displacement of one or both of the eyes' lenses. The most serious complications affect the cardiovascular system, especially heart valves and the aorta.*

(Credit: Koller 2016)

Lexical Semantics in Computational Linguistics

- Many words are *synonymous*, or at least *semantically similar*
- *He has passed on, met his maker, kicked the bucket, expired, ceased to be!*

Information Retrieval

- Goal to find relevant documents, even if differently phrased
- QUERY: “female astronauts”
- DOCUMENT: “In the history of the Soviet space program, there were only three female cosmonauts: Valentina Tereshkova, Svetlana Savitskaya, and Elena Kondakova”
- System must recognize that *astronaut* and *cosmonaut* have similar meanings (in a given context!).

(Credit: Koller 2016)

Machine Translation

The box is in the pen. Bar-Hillel (1960)

- World knowledge necessary to disambiguate *polysemous* words
- Correct translation depends on selecting the correct sense of *pen*

(Credit: Koller 2016)

(Back to) Classical Lexical Semantics

- **Polysemy**: Word has two different meanings that are clearly related to each other
  - School_1: institution at which students learn
  - School_2: building that houses school_1
- **Homonyny**: Word has two different meanings that have no obvious relation to each other.
  - Bank_1: financial institution
  - Bank_2: land alongside a body of water

(Credit: Koller 2016)

Word Sense Disambiguation

- Word sense disambiguation is the problem of tagging each word token with its word sense.
- WSD accuracy depends on sense inventory; state of the art is above 90% on coarse-grained senses
- Techniques tend to combine supervised training on small amount of annotated data with unsupervised methods.

(Credit: Koller 2016)

Problem

- Hand-written thesauruses much too small
  - English Wordnet: 117.000 synsets
  - GermaNet: 85.000 synsets
- Number of word types in English Google n-gram corpus: > 1 million.
- This is not how we can solve the query expansion problem
- Can we learn lexical semantic knowledge automatically?
  - …and in a way that is cognitively sound?

Meaning and Distribution

We found a little, hairy wampimuk sleeping behind the tree.

Meaning and Distribution

We found a little, hairy wampimuk sleeping behind the tree.

- "Die Bedeutung eines Wortes liegt in seinem Gebrauch." (Ludwig Wittgenstein)
  - meaning = use = distribution in language

Meaning and Distribution

We found a little, hairy wampimuk sleeping behind the tree.

- "Die Bedeutung eines Wortes liegt in seinem Gebrauch." (Ludwig Wittgenstein)
  - meaning = use = distribution in language
- "You shall know a word by the company it keeps." (Firth, 1957)
  - distribution = collocations = habitual word combinations

Meaning and Distribution

We found a little, hairy wampimuk sleeping behind the tree.

- “Die Bedeutung eines Wortes liegt in seinem Gebrauch.” (Ludwig Wittgenstein)
  - meaning = use = distribution in language
- “You shall know a word by the company it keeps.” (Firth, 1957)
  - distribution = collocations = habitual word combinations
- Distributional hypothesis: difference of meaning correlates with difference of distribution (Zellig Harris, 1954)
  - semantic distance
  - Assumption: Semantically similar words tend to occur in the context of the same words. $\longrightarrow$ “similar” as approximation of “synonymous”

Meaning and Distribution

We found a little, hairy wampimuk sleeping behind the tree.

- "Die Bedeutung eines Wortes liegt in seinem Gebrauch." (Ludwig Wittgenstein)
  - meaning = use = distribution in language
- "You shall know a word by the company it keeps." (Firth, 1957)
  - distribution = collocations = habitual word combinations
- Distributional hypothesis: difference of meaning correlates with difference of distribution (Zellig Harris, 1954)
  - semantic distance
  - Assumption: Semantically similar words tend to occur in the context of the same words. ⟶ "similar" as approximation of "synonymous"
- "What people know when they say that they know a word is not how to recite its dictionary definition – they know how to use it [...] in everyday discourse." (Miller, 1986)

What does “bardiwac” mean?

- *He handed her a glass of bardiwacs.*
- *Beef dishes are made to complement the bardiwacs.*
- *Nigel staggered to his feet, face flushed from too much bardiwac.*
- *Malbec, one of the lesser-known bardiwac grapes, responds well to Australia’s sunshine.*
- *I dined off bread and cheese and this excellent bardiwac.*
- *The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.*

(Stefan Evert, tutorial at NAACL 2010)

What does "bardiwac" mean?

- *He handed her a glass of bardiwacs.*
- *Beef dishes are made to complement the bardiwacs.*
- *Nigel staggered to his feet, face flushed from too much bardiwac.*
- *Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.*
- *I dined off bread and cheese and this excellent bardiwac.*
- *The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.*

⟶ Bardiwac is a red wine

Distributional semantics

Landauer and Dumais 1997, Turney and Pantel 2010, ...

```
he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
 surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
 in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
 But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
 they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
 the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind
```

Distributional semantics

The geometry of meaning

Distributional Semantic Model (DSM): a scaled and/or transformed co-occurrence matrix $\mathbf{M}$, such that each row $\mathbf{x}$ represents the distribution of a target term across contexts.

- e.g., within a document, within a window of [content] words before and after, etc.

	shadow	shine	planet	night
moon	16	29	10	22
sun	15	45	14	10
dog	10	0	0	4

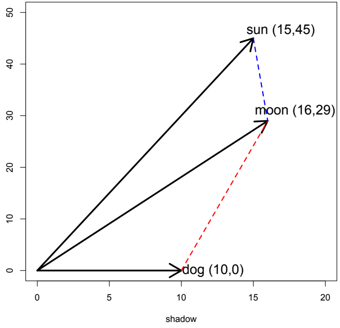
Distributional semantics

The geometry of meaning

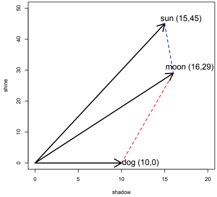
Distributional Semantic Model (DSM): a scaled and/or transformed co-occurrence matrix $\mathbf{M}$, such that each row $\mathbf{x}$ represents the distribution of a target term across contexts.

- e.g., within a document, within a window of [content] words before and after, etc.

	shadow	shine	planet	night
moon	16	29	10	22
sun	15	45	14	10
dog	10	0	0	4

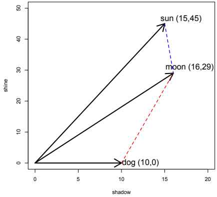


Lexical similarity



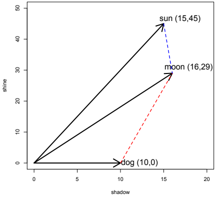
- Semantic similarity approximated by geometric distance of vectors (angle)
  - (correctly) ignores length of vectors (= frequency of words)
  - similar angle = similar proportion of context words

Lexical similarity



- Semantic similarity approximated by geometric distance of vectors (angle)
  - (correctly) ignores length of vectors (= frequency of words)
  - similar angle = similar proportion of context words
- Cosine of an angle is easy to compute
  - cos $\longrightarrow$ 1: angle is 0° (very similar)
  - cos $\longrightarrow$ 0: angle is 90° (very dissimilar)

Lexical similarity



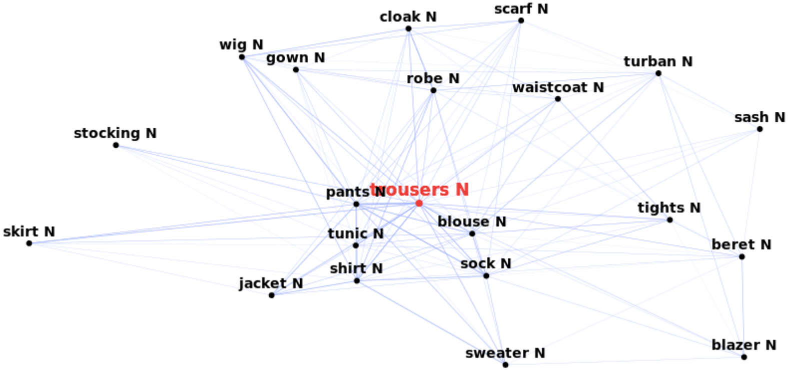
- Semantic similarity approximated by geometric distance of vectors (angle)
  - (correctly) ignores length of vectors (= frequency of words)
  - similar angle = similar proportion of context words
- Cosine of an angle is easy to compute
  - cos $\longrightarrow$ 1: angle is 0° (very similar)
  - cos $\longrightarrow$ 0: angle is 90° (very dissimilar)
- successful in tasks that concern content words: detecting synonyms, lexical entailment, ...
  - see Turney & Pantel, 2010; Baroni & Lenci, 2010, among others

Distributional Semantic Models

	get	see	use	hear	eat	kill
knife	0.027	-0.024	0.206	-0.022	-0.044	-0.042
cat	0.031	0.143	-0.243	-0.015	-0.009	0.131
dog	-0.026	0.021	-0.212	0.064	0.013	0.014
boat	-0.022	0.009	-0.044	-0.040	-0.074	-0.042
cup	-0.014	-0.173	-0.249	-0.099	-0.119	-0.042
pig	-0.069	0.094	-0.158	0.000	0.094	0.265
banana	0.047	-0.139	-0.104	-0.022	0.267	-0.042

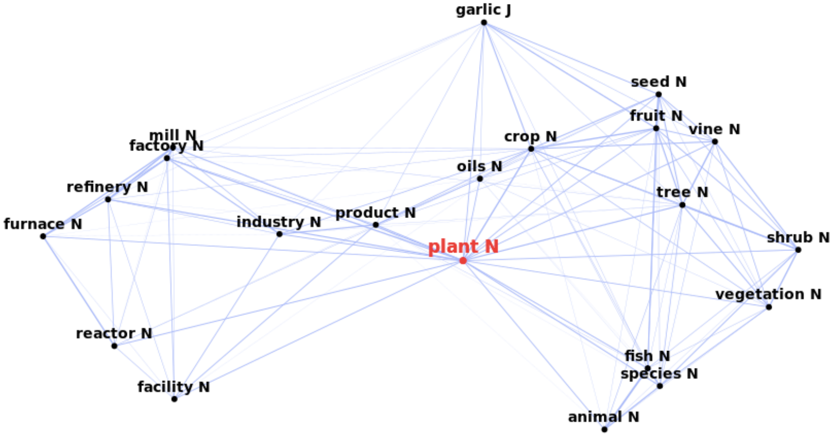
(Evert et al, tutorial at ESSLLI 2018)

Nearest Neighbors of *trousers*



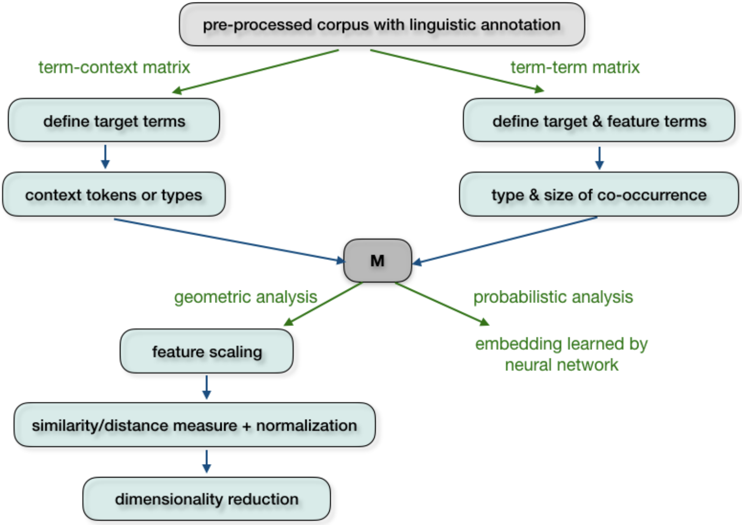
*Based on DSM built on EN Wikipedia, (filtered) dependency contexts

Nearest Neighbors of *trousers*



*Based on DSM built on EN Wikipedia, (filtered) dependency contexts

Building a distributional model



(Evert et al, tutorial at ESSLLI 2018)

Linguistic Preprocessing

Defining a term

- Tokenization
- POS-tagging (*light*_N vs. *light*_J vs. *light*_V)
- Stemming/lemmatization
  - *go*, *goes*, *went*, *gone*, *going* → *go*
- Dependency parsing or shallow syntactic chunking

(Credit: Teufel 2017)

Linguistic Preprocessing

Defining a term

- Tokenization
- POS-tagging (*light*_N vs. *light*_J vs. *light*_V)
- Stemming/lemmatization
  - *go*, *goes*, *went*, *gone*, *going* → *go*
- Dependency parsing or shallow syntactic chunking

Effect of linguistic preprocessing

- Nearest neighbors of *walk* (BNC, DSM defined by head of the subject of *walk*)
  - **Word forms**: stroll, walking, walked, go, path, drive, ride, wander, sprinted, sauntered
  - **Lemmatized forms**: hurry, stroll, stride, trudge, amble, wander, walk-NN, walking, retrace, scuttle

(Credit: Teufel 2017)

Term-document vs. term-term matrices

- In IR, the "context" is always exactly one document
- This results in term-document matrices (aka "Vector Space Models")
- This allows us to measure the similarity of words with sets of words (e.g. documents vs. queries in IR)
- Term-document matrices are sparse

Context Type

- Context term appears in same fixed **window**
- Context term is a member in the same **linguistic unit** as target (e.g. paragraph, sentence, turn in conversation)
- Context term is linked to target by a **syntactic dependency** (e.g. subject, modifier)

(Credit: Teufel 2017)

Context Type

- Context term appears in same fixed **window**
- Context term is a member in the same **linguistic unit** as target (e.g. paragraph, sentence, turn in conversation)
- Context term is linked to target by a **syntactic dependency** (e.g. subject, modifier)
- Context type (e.g. window size) can have impact on how terms are related to those in its nearest neighborhood
  - For example, the tendency for smaller window sizes is to be pragmatically related (e.g. car, van, vehicle, truck), while in larger window sizes syntagmatically related (e.g. car, drive, park, windscreen)

(Credit: Teufel 2017)

Similarity vs. Relatedness

It is generally accepted that there are (at least) two dimensions of word associations:

- **Semantic Similarity**: two words sharing a high number of salient features (attributes) $\longrightarrow$ *paradigmatic relatedness*
  - (near) synonymy (*car-automobile*)
  - hyperonymy (*car-vehicle*)
  - co-hyponymy (*car-van-lorry-bike*)

Similarity vs. Relatedness

It is generally accepted that there are (at least) two dimensions of word associations:

- **Semantic Similarity**: two words sharing a high number of salient features (attributes) ⟶ *paradigmatic relatedness*
  - (near) synonymy (*car-automobile*)
  - hyperonymy (*car-vehicle*)
  - co-hyponymy (*car-van-lorry-bike*)
- **Semantic Relatedness**: two words semantically associated without being necessarily similar ⟶ *syntagmatic relatedness*
  - function (*car-drive*)
  - meronymy (*car-tire*)
  - location (*car-road*)
  - attribute (*car-fast*)
  - other (*car-petrol*)

(Credit: Teufel 2017)

Feature Scaling

Feature scaling is used to “discount” less important features:

- Logarithmic scaling: $O' = log(O + 1)$ (cf. Weber-Fechner law for human perception)

Feature Scaling

Feature scaling is used to "discount" less important features:

- Logarithmic scaling: $O' = log(O + 1)$ (cf. Weber-Fechner law for human perception)
- Relevance weighting, e.g. tf.idf (information retrieval)
  - $tf.idf = tf \cdot log(D/df)$
  - tf = co-occurrence frequency O
  - df = document frequency of feature (or nonzero count)
  - D = total number of documents (or row count of $\mathbf{M}$

Feature Scaling

Feature scaling is used to "discount" less important features:

- Logarithmic scaling: $O' = log(O + 1)$ (cf. Weber-Fechner law for human perception)
- Relevance weighting, e.g. tf.idf (information retrieval)
  - $tf.idf = tf \cdot log(D/df)$
  - tf = co-occurrence frequency O
  - df = document frequency of feature (or nonzero count)
  - D = total number of documents (or row count of $\mathbf{M}$
- Statistical **association measures** (Evert 2004, 2008) take frequency of target term and feature into account
  - often based on comparison of observed and expected co-occurence frequency (how surprised are we to see context term associated with target word?)
  - measures differ in how they balance O and E

Simple association measures

- **Pointwise Mutual Information** (PMI): compares observed vs. expected frequency of a word combination

$$PMI(w_1, w_2) = log_2 \frac{f_{obs}}{f_{exp}}$$

  - Disadvantage: PMI overrates combinations involving rare terms

Simple association measures

- **Pointwise Mutual Information** (PMI): compares observed vs. expected frequency of a word combination

$$PMI(w_1, w_2) = log_2 \frac{f_{obs}}{f_{exp}}$$

  - Disadvantage: PMI overrates combinations involving rare terms

- **t-score**: How many standard deviations is f_{obs} away from assumed mean (f_{exp})?

$$assoc_{t-test}(w_1, w_2) = \frac{f_{obs} - f_{exp}}{\sqrt{f_{obs}}}$$

Simple association measures

- **Pointwise Mutual Information** (PMI): compares observed vs. expected frequency of a word combination

$$PMI(w_1, w_2) = log_2 \frac{f_{obs}}{f_{exp}}$$

  - Disadvantage: PMI overrates combinations involving rare terms

- **t-score**: How many standard deviations is f_{obs} away from assumed mean (f_{exp})?

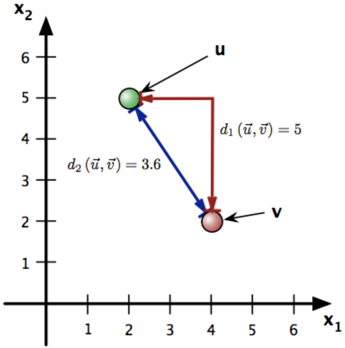
$$assoc_{t-test}(w_1, w_2) = \frac{f_{obs} - f_{exp}}{\sqrt{f_{obs}}}$$

- **Log-Likelihood** (Dunning, 1993): describes relative probability of obtaining the observed frequency for all permissible values of the parameters

$$G^2 = \pm 2 \cdot \left(f_{obs} \cdot log_2 \frac{f_{obs}}{f_{exp}} - (f_{obs} - f_{exp}) \right)$$

Geometric Distance

- **Distance** between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n \rightarrow$ (dis)similarity
  - $\mathbf{u} = (u_1, \ldots, u_n)$
  - $\mathbf{v} = (v_1, \ldots, v_n)$
- **Euclidean** distance $d_2(\mathbf{u}, \mathbf{v})$
- "City block" **Manhattan** distance $d_1(\mathbf{u}, \mathbf{v})$
- Both are special caes of the **Minkowski** p-distance $d_p(\mathbf{u}, \mathbf{v})$ (for $p \in [1, \infty]$)



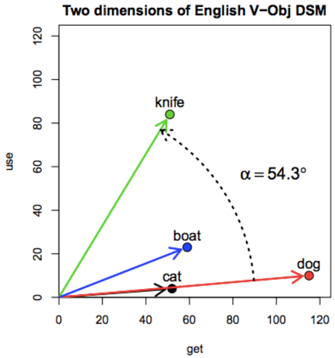
(Evert et al, tutorial at ESSLLI 2018)

Similarity Measures

- Angle α between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ is given by

$$cos\alpha = \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}$$

- **Cosine** measure of similarity: $cos\alpha$
  - $cos\alpha = 1 \rightarrow$ collinear
  - $cos\alpha = 0 \rightarrow$ orthogonal
- Corresponding metric: **angular distance** α

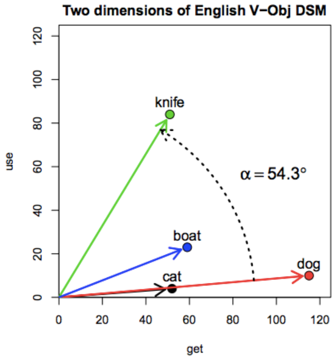


Similarity Measures

- Angle α between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ is given by

$$cos\alpha = \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}$$

- **Cosine** measure of similarity: $cos\alpha$
  - $cos\alpha = 1 \rightarrow$ collinear
  - $cos\alpha = 0 \rightarrow$ orthogonal
- Corresponding metric: **angular distance** α



Euclidean distance or cosine similarity?

- They are the equivalent: if vectors have been normalized ($\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$), both lead to the same neighborhood ranking.

(Evert et al, tutorial at ESSLLI 2018)

LSA

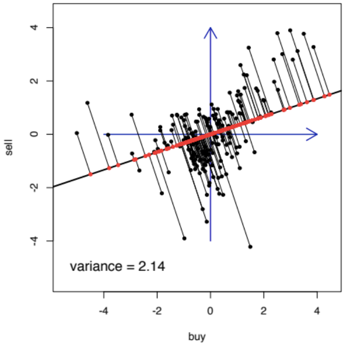
- Vectors in standard vector space are very sparse
- Orthogonal dimensions clearly wrong for near-synonyms *canine-dog*
- Different word senses are conflated into the same dimension
- One way to solve this: **dimensionality reduction**
- Hypothesis for LSA (Latent Semantic Analysis; Landauer): true semantic space has fewer dimensions than number of words observed
- Extra dimensions are noise. Dropping them brings out **latent** semantic space

Dimensionality reduction by PCA

- Principal component analysis (**PCA**)
  - orthogonal projection into orthogonal latent dimensions
  - finds optimal subjspace of given dimensionality (such that orthogonal projection preserves distance information)
  - but requires centered features → no longer sparse
- Singular value decomposition (**SVD**)
  - the mathematical algorithm behind PCA
  - often applied without centering in distributional semantics
  - note: optimality of subspace no guaranteed
- NB: row vectors should be re-normalized after PCA/SVD
  - unless cosine similarity / angular distance is used
  - also normalize vectors **before** dimensionality reduction

Dimensionality reduction by RI

- Random indexing (**RI**)
  - Project into random subspace (Sahlgren & Karlgren, 2005)
  - reasonably good if there are many subspace dimensions
  - can be performed online without collecting full co-occurrence matrix



Some applications in computational linguistics

- Query expansion in IR (Grefenstette, 1994)
- Unsupervised POS induction (Schütze, 1995)
- Word sense disambiguation (Schütze, 1998; Rapp, 2004)
- Thesuarus compilation (Lin 1998; Rapp 2004)
- Attachment disambiguation (Pantel & Lin, 2000)
- Probabilistic language models (Bengio et al, 2003)
- Translation equivalents (Sahlgren & Karlgren, 2005)
- Ontology & wordnet expansion (Pantel et al, 2009)
- Language change (Sagi et al, 2009; Hamilton et al, 2016)
- Multiword expressions (Kiela & Clark, 2013)
- Analogies (Turney 2013; Gladkova et al, 2016)
- Sentiment analysis (Rothe & Schütze, 2016; Yu et al, 2017)
- $\longrightarrow$ Input representations for neural networks & machine learning

Software packages

Infomap NLP	C	*classical LSA-style DSM*
HiDEx	C++	*re-implementation of the HAL model (Lund & Burgess, 1996)*
SemanticVectors	Java	*scalable architecture based on random indexing representation*
S-Space	Java	*complex object-oriented framework*
JoBimText	Java	*UIMA / Hadoop framework*
Gensim	Python	*complex framework, focus on parallelization and out-of-core algorithms*
Vecto	Python	*framework for count & predict models*
DISSECT	Python	*user-friendly, designed for research on compositional semantics*
wordspace	R	*interactive research laboratory, but scales to real-life data sets*

Assignment 2

- Assignment posted on course website, **Due date: June 5th**
- You will implement a DSM using the `wordspace` package in `R`
- Software installation:
  - `R` version 3.5 or newer from `http://www.r-project.org/`
  - `R` packages from CRAN (through RStudio menu): `sparsesvd`, `wordspace`
  - Get data sets, precompiled DSMs and `wordspaceEval` from `http://wordspace.collocations.de/doku.php/course:material`
- You can also explore some DSM similarity networks online:
  - `https://corpora.linguistik.uni-erlangen.de/shiny/wordspace/`
  - built in `R` with `wordspace` and `shiny`