Chapter NLP:VIII

VIII. Text Representation Models

- Introduction to Text Representation Models
- Bag of Words / Vector Space Model
- Similarity Measures in Natural Language Processing

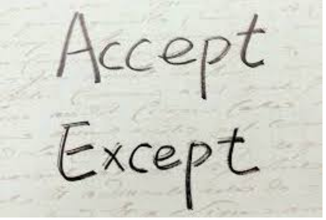
Similarity Measures

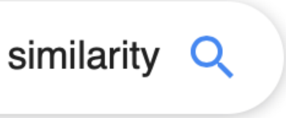
- A similarity measure is a real-valued function that quantifies how similar two instances of the same concept are.
- Usually, possible values range between 0 (no similarity) and 1 (identity).
- In NLP, instances are (the representations of) input text spans.

Various use cases in NLP

- Clustering
- Spelling correction
- Retrieval of relevant web pages
- Detection of related documents
- Paraphrase recognition
- (Near-) Duplicate or text reuse detection
- Identification of counterarguments

. . . and many more





Similarity Measures

Text Similarity

- Similarity between the *form* of two texts or text spans.
- Similarity between the *meaning* of two texts or text spans.

  Similar form, different meaning: “This is shit.” vs. “This is *the* shit.”

  Other way round: “Obama visited the capital of France.” vs. “Barack Obama was in Paris.”
- Ultimately, similarity measures aim to capture the latter.
- But the former is often used as a proxy.

Text similarity measures

- Vector-based measures. Mainly, for similarities between feature vectors.
- Edit distance. For spelling similarities.
- Thesaurus methods. For synonymy-related similarities.
- Distributional similarity. For similarities in the contextual usage.

  Clustering is mostly based on the first, but the others may still be used internally.

Similarity Measures

Vector-based Similarity Measures

- Given a collection of input texts or text spans, the goal is to compare any two instances o_1, o_2 from them.
- Comparison is done on feature-based representations (i.e., o_1 and o_2 are mapped to feature vectors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, respectively).

Feature-based representation (recap)

- A feature vector is an ordered set of values of the form $\mathbf{x} = (x_1, \ldots, x_m)$, where each feature x_i denotes a measurable property of an input.

  We consider only real-valued features here.

- Each instance o_j is mapped to a vector $\mathbf{x}^{(j)} = (x_1^{(j)}, \ldots, x_m^{(j)})$ where $x_i^{(j)}$ denotes the value of feature x_i for o_j.

  We consider only values normalized to the range $[0, 1]$ here.

Similarity measures and clustering

- Clustering mostly relies on vector-based similarity measures.

Similarity Measures

Vector-based Similarity Measures: Concept

Measuring similarity between vectors

- Compare two vectors of the same representation with each other.

  $(1.0, 0.0, 0.3)$ vs. $(0.0, 0.0, 0.7)$ for $\mathbf{x} = ($red, green, blue$)$

- The difference of each vector dimension is computed individually.

  1.0 vs. 0.0 0.0 vs. 0.0 0.3 vs. 0.7

- The similarity results from an aggregation of all differences.

  For example: $\frac{1.0+0.0+0.4}{3} \approx 0.467$

Concrete similarity measures

- Numerous vector-based measures are found in the literature [Cha, 2007].
- We focus on four of the most common measures here: *Cosine similarity*, *Jaccard similarity*, *Euclidean distance*, and *Manhattan distance*.

  As mentioned before, distance can be seen as the inverse of similarity.

Similarity Measures

Vector-based Similarity Measures: Distance Functions

Properties of a distance function (aka metric)

- Non-negativity. $d(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \geq 0$
- Identity. $d(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) = 0$
- Symmetry. $d(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = d(\mathbf{x}^{(2)}, \mathbf{x}^{(1)})$
- Subadditivity. $d(\mathbf{x}^{(1)}, \mathbf{x}^{(3)}) \leq d(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) + d(\mathbf{x}^{(2)}, \mathbf{x}^{(3)})$

  Clustering actually does not necessarily require subadditivity.

Distance computation in clustering

- Internally, clustering algorithms compute distances between instances.

	x_1	x_2	$\ldots$	x_m
$\mathbf{x}^{(1)}$	$x_1^{(1)}$	$x_2^{(1)}$	$\ldots$	$x_m^{(1)}$
$\mathbf{x}^{(2)}$	$x_1^{(2)}$	$x_2^{(2)}$	$\ldots$	$x_m^{(2)}$
$\vdots$				
$\mathbf{x}^{(n)}$	$x_1^{(n)}$	$x_2^{(n)}$	$\ldots$	$x_m^{(n)}$

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\ldots$	$\mathbf{x}^{(n)}$
$\mathbf{x}^{(1)}$	0	$d(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$	$\ldots$	$d(\mathbf{x}^{(1)}, \mathbf{x}^{(n)})$
$\mathbf{x}^{(2)}$	-	0	$\ldots$	$d(\mathbf{x}^{(2)}, \mathbf{x}^{(n)})$
$\vdots$				
$\mathbf{x}^{(n)}$	-	-	$\ldots$	0

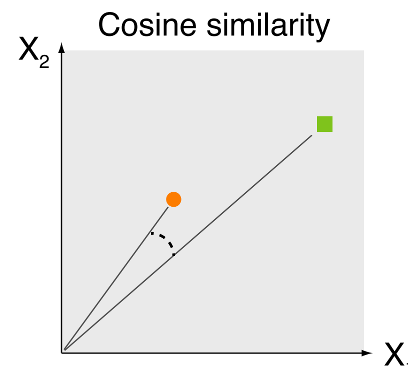
Similarity Measures

Vector-based Similarity Measures: Cosine Similarity

Cosine similarity (aka cosine score)

- Cosine similarity captures the cosine of the angle between two feature vectors.
- The smaller the angle, the more similar the vectors.

  This works because cosine is maximal for 0°.
- $||\mathbf{x}||$ denotes the L2 norm of vector $\mathbf{x}$:



$$sim_{Cosine}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{\mathbf{x}^{(1)} \cdot \mathbf{x}^{(2)}}{||\mathbf{x}^{(1)}|| \cdot ||\mathbf{x}^{(2)}||} = \frac{\sum_{i=1}^{m} x_i^{(1)} \cdot x_i^{(2)}}{\sqrt{\sum_{i=1}^{m} x_i^{(1)^2}} \cdot \sqrt{\sum_{i=1}^{m} x_i^{(2)^2}}}$$

Notice

- The cosine similarity abstracts from the length of the vectors.
- Angle computation works for any number of dimensions.
- Cosine similarity is the most common similarity measure.

Similarity Measures

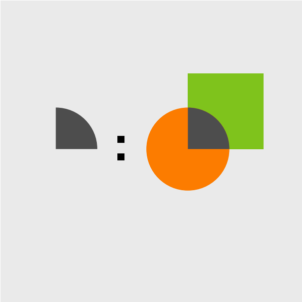
Vector-based Similarity Measures: Jaccard Similarity

Jaccard similarity coefficient (aka Jaccard index)

- The Jaccard coefficient captures how large the intersection of two sets is compared to their union.
- With respect to vector representations, this makes at least sense for Boolean features.

  For others, if there is a reasonable way of thresholding.

Jaccard similarity



$$sim_{Jaccard}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{|\mathbf{x}^{(1)} \cap \mathbf{x}^{(2)}|}{|\mathbf{x}^{(1)} \cup \mathbf{x}^{(2)}|} = \frac{|\mathbf{x}^{(1)} \cap \mathbf{x}^{(2)}|}{|\mathbf{x}^{(1)}| + |\mathbf{x}^{(2)}| - |\mathbf{x}^{(1)} \cap \mathbf{x}^{(2)}|}$$

$$= \frac{\sum_{x_i^{(1)} = x_i^{(2)}} 1}{m + m - \sum_{x_i^{(1)} = x_i^{(2)}} 1}$$

Notice

- The Jaccard similarity does *not* consider the size of the difference between feature values.

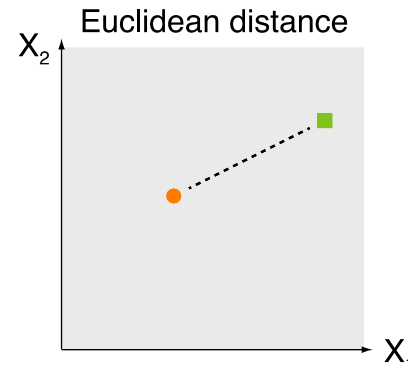
Similarity Measures

Vector-based Similarity Measures: Euclidean Similarity

Euclidean distance

- The Euclidean distance captures the absolute straight-line distance between two feature vectors.

$$\mathit{dist}_{\mathit{Euclidean}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sqrt{\sum_{i=1}^{m} |x_i^{(1)} - x_i^{(2)}|^2}$$



Euclidean similarity

- If all feature values are normalized to $[0, 1]$, the Euclidean similarity is:

$$\mathit{sim}_{\mathit{Euclidean}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = 1 - \frac{\mathit{dist}_{\mathit{Euclidean}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})}{\sqrt{m}}$$

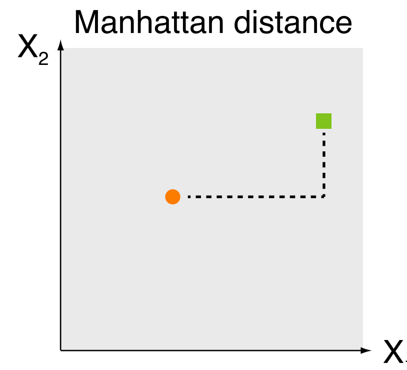
Notice

- Euclidean spaces generalize to any number of dimensions $m \geq 1$.
- Here, this means to any number of features.

Similarity Measures

Vector-based Similarity Measures: Manhattan Similarity

Manhattan distance (aka city block distance)



- The Manhattan distance is the sum of all absolute differences between two feature vectors.

$$\mathit{dist}_{\mathit{Manhattan}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \;=\; \sum_{i=1}^{m} |x_i^{(1)} - x_i^{(2)}|$$

Manhattan similarity

- If all feature values are normalized to $[0, 1]$, the Manhattan similarity is:

$$\mathit{sim}_{\mathit{Manhattan}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \;=\; 1 - \frac{\mathit{dist}_{\mathit{Manhattan}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})}{m}$$

Notice

- Manhattan distance and Euclidean distance are both special cases of the *Minkowski distance*.

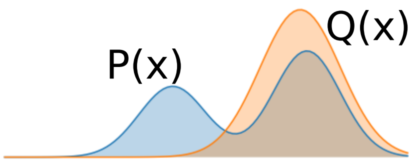
$$\mathit{dist}_{\mathit{Minkowski}}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \sqrt[p]{\sum_{i=1}^{m} |\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}|^p} \quad \text{for any } p \in \mathbb{N}^+$$

Similarity Measures

Vector-based Similarity Measures: Kullback-Leibler-Divergence, Jenson-Shannon-Divergence

Kullback–Leibler–Divergence (KL)

- A measure of how one probability distribution is different from a second in terms of information gain (asymmetric measure, does not qualify as a statistical metric of spread - it also does not satisfy the triangle inequality)



$$D_{\mathsf{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

Jenson-Shannon-Divergence (JSD)

- JSD is based on the Kullback–Leibler divergence, with some notable (and useful) differences, including that it is symmetric and it always has a finite value.

$$\begin{aligned} sim_{\mathsf{JSD}}(P(x) \parallel Q(x)) &= 1 - \left(\frac{1}{2} D_{\mathsf{KL}}(P(x) \parallel M(x)) + \frac{1}{2} D_{\mathsf{KL}}(Q(x) \parallel M(x)) \right) \\ M(x) &= \frac{1}{2}(P(x) + Q(x)) \end{aligned}$$

Notice

- This kind of distances are used in probability mathematical spaces which are not linear (e.g. Multinomial Distributions in Topic Models)

Similarity Measures

Vector-based Similarity Measures: When to Use What Measure?

Comparison of the measures

- Cosine similarity. Puts the focus on those properties that occur. Targets situations where a vector's direction matters rather than its length.
  A prominent use case is matching queries with documents in web search.
- Jaccard similarity. Seems less precise than cosine similarity, but this also makes it more robust (it "overfits" less).
- Euclidean and Manhattan. Target situations where a value of 0 does not mean the absence of a property.
- Euclidean or Manhattan. Depends on whether sensitivity to outliers in certain dimensions is preferred or not.
- Jenson–Shannon. If the text representation is expressed in terms of probability distributions.

Similarity as an optimization hyperparameter

- In general, it is not always clear what measure will prove best.
- One way to deal with this is to simply evaluate different measures.
- In some applications, all measures can be used simultaneously.

Similarity Measures

Similarity between Strings

Limitation of vector-based measures in NLP

- Similarity is defined based on corresponding feature values $x_j^{(1)}, x_j^{(2)}$.
- Most features in NLP are derived directly from text spans.
- Similarity between different forms with similar meaning is missed ...

  "traveling" vs. "travelling"   "woodchuck" vs. "groundhog"   "Trump" vs. "The President"

- ... unless such differences are accounted for.

Similar strings

- May contain differences in writing, due to spelling errors, language variations, or additional words.
- May contain different words that refer to similar concepts.
- May contain different concepts that are related in a way that should be seen as similar in a given application.

  ... and similar

Similarity Measures

Similarity between Strings: Edit Distance

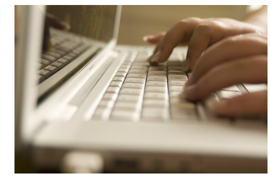
What is (minimum) edit distance?

❑ The minimum number (or cost) of editing operations needed to transform one string to another.

❑ Editing operations. Insertion, deletion, substitution.

❑ Weighted edit distance. Different edits vary in costs.

```
I N T E * N T I O N
| | | | | | | | | |
d s s   i s
| | | | | | | | | |
* E X E C U T I O N
```



How to compute edit distance?

❑ Sequence alignment using dynamic programming.

❑ Equals shortest path search in a weighted graph.

	E	X	E
I	s(I,E)	i(*,X)	
N	d(N,*)	s(N,X)	
T			

Selected applications

❑ Spelling correction (e.g., web search queries).

"westauwang" $\rightarrow$ Did you mean "restaurant"?

❑ Alignment in computational biology (kind of a language problem).

Similarity Measures

Similarity between Strings: Thesaurus Methods

What are synonyms?

- Words (or terms) that have the same meaning in some or all contexts.

  "couch" vs. "sofa" "big" vs. "large" "water" vs. "$H_2 0$" "vomit" vs. "throw up"

- There are hardly any perfectly synonymous terms.

  Even seemingly identical terms usually differ in terms of politeness, slang, genre, etc.

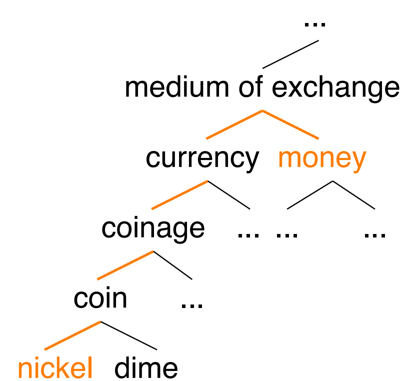
- Synonymy is a relation between senses rather than words.

  "big" vs. "large" $\rightarrow$ "Max became kind of a <insert> brother to Linda."

How to identify related senses?

- Compute distance in thesauri, such as *WordNet*.
  wordnetweb.princeton.edu/perl/webwn

  S: (n) **nickel**, Ni, atomic number 28 (a hard malleable ductile silvery metallic element that is resistant to corrosion; used in alloys; occurs in pentlandite and smaltite and garnierite and millerite)
  S: (n) **nickel** (a United States coin worth one twentieth of a dollar)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) coin (a flat metal piece (usually a disc) used as money)



- Several libraries for such measures freely available.

Similarity Measures

Similarity between Strings: Distributional Similarity

Limitation of thesaurus methods

- Many words are missing as well as basically all phrases, and also some sense connections.
- Verbs and adjectives are not as hierarchically structured as nouns.
- Thesauri are not available for all languages.

Idea of distributional similarity

"You shall know a word by the company it keeps!" [Firth, 1957]

- If A and B have almost identical environments, they are synonyms.
- Two words are similar if they have similar word contexts (i.e., if they have similar words around them).

"Everybody likes tesgüino." "A bottle of tesgüino is on the table."
"Tesgüino makes you drunk." "We make tesgüino out of corn."

→ An alcoholic beverage like beer.

Similarity Measures

Similarity between Strings: Distributional Hypothesis

Word-context matrix

- Co-occurrences of words in a corpus within a window of some number of words (say, 20).

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

Distributional Similarity between words

- Term (Word)–Context–Matrix can be used to calculate semantic similarity based on Cosine Similarity, Pointwise Mutual Information or JSD
- See section **Distributional Hypothesis** and **Cooccurrence Analysis** for details

Similarity Measures

Similarity between Strings: From Strings back to Texts

Encoding similarities in feature vectors

- String similarities can be used in diverse ways within features.

  Frequency of ~~“money”~~ the sense “the most common medium of exchange”

  Frequency of all writings of “traveling”

- Where reasonable, embeddings can simply be used as feature vectors.

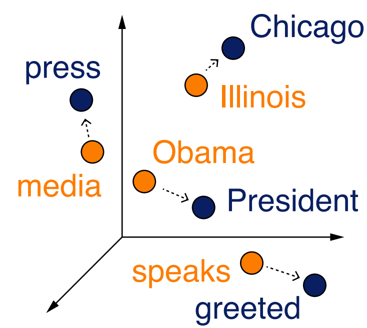
  “nickel” $\rightarrow$ (0.14, 0.03, 0.44, . . . , 0.22)   “money” $\rightarrow$ (0.18, 0.06, 0.49, . . . , 0.01)

Word Mover’s Distance [Kusner et al., 2015]

- The distance of the optimal alignment of two texts.

  Obama speaks to the media in Illinois

  The press is greeted by the President in Chicago



- Represents texts by sequences of word embeddings.