NLP
Text Similarity

Text Kernels
Need for Kernels

• Sentences (or documents, or words) can be represented as vectors in many different ways, depending on the type of similarity that needs to be computed.

• In document retrieval, word order may often be ignored (the so-called “bag of words” approach).
  – For example, “… cat … dog” may be considered the same as “… dog … cat.”

• In sentence retrieval word order may matter,
  – e.g., compare “Utopia invades Asgård” and “Asgård invades Utopia.”
  – A bag of words approach, however, will treat these two sentences as the same.
The Basic Idea behind Kernels
The Basic Idea behind Kernels
The Basic Idea behind Kernels
The Basic Idea behind Kernels

\[ \varphi(x) \]
Unigram Kernels

• A unigram kernel is essentially the same as a bag of words
• For example, we can split the sentence “Utopia invades Asgård” into the three unigrams:
  – “Utopia”
  – “invades”
  – “Asgård”
• Similarly, we can split “Asgård invades Utopia” into:
  – “Asgård”
  – “invades”
  – “Utopia”
• However, the vector representations of the two sentences will be identical (<1,1,1>) and therefore their cosine similarity will be computed as 1.
Bigram Kernels

• Using a bigram kernel, we can split the sentence “Utopia invades Asgård” into the two bigrams:
  – “Utopia invades”
  – “invades Asgård”

• Similarly, we can split “Asgård invades Utopia” into:
  – “Asgård invades”
  – “invades Utopia”

• Using this representation, the vectors corresponding to the two sentences are \( <1,1,0,0> \) and \( <0,0,1,1> \), respectively, and their cosine similarity is 0.
Longer n-grams

• N-grams for n>2 can also be used for measuring similarity.
• Longer n-grams (shingles) are particularly useful in Information Retrieval, for fast retrieval of similar documents.
• In NLP, people use n-gram similarity methods for the evaluation of
  – machine translation (BLEU – Papineni et al.)
  – text summarization (ROUGE – Lin et al.)
• In both cases, an automatically generated text (a translation or a summary) is compared against a set of reference sentences.
One hundred artists from 16 countries will exhibit 270 pieces of **artistic work made on porcelain** in the Palace of Arts in Cairo in an exhibition inaugurated earlier this week and that will last for two weeks.

One hundred artists from 61 states are exhibiting 270 pieces of **porcelain artwork** at the Arts Palace in Cairo, as part of a two-week exhibition that opened at the beginning of this week.

100 artists from 16 countries are exhibiting 270 pieces of work on **porcelain** in Cairo's Arts Palace as part of a 2-week exhibition, which opened at the beginning of the week.

One hundred artists from 16 countries will display 270 pieces of art works **undertaken on porcelain** in the Palace of Arts in Cairo in an exhibition opened at the beginning of this week and which will continue for two weeks.

A hundred artists from 16 countries will exhibit 270 pieces of artistic works **that are made on Porcelain** in the arts palace in Cairo in an exhibition inaugurated earlier this week, which will last for two weeks.

One hundred artists from 61 countries exhibited 270 types of art pieces **done on porcelain** in the Cultural Palace of Cairo, in an exhibition that was inaugurated at the beginning of the current week. The exhibition would last two weeks.

One hundred artists from 16 countries will exhibit 270 works of art **made with porcelain** in the Arts Castle in Cairo in an exhibition that opened earlier this week and will last for two weeks.

A hundred artists from 16 countries are displaying 270 **porcelain** art pieces in an exhibition that opened early in the week, and will continue over a span of two weeks.

A hundred artists from 16 countries display 270 artistic **porcelain** pieces at the Palace of Arts in Cairo, in an exhibition that was opened earlier this week and continuing for two weeks.
Letter and Substring Kernels

- Letter n-grams can be used for various applications such as spelling correction, language recognition, and named entity recognition.
- For example, the word *stop* can be represented as the set of all of its substrings: s, t, o, p, st, to, op, sto, top, and stop.
- In this representation, $\text{sim}(\text{stop, stops}) > \text{sim}(\text{stop, plot})$, even though all three words are different.
Subsequences

• Unlike a substring, a subsequence doesn’t need to consist of contiguous words (or letters).
• *comp, cotr, opter, cpute* – all of these are letter-based subsequences of *computer*
• Subsequence kernels (of words, not letters) are most useful for measuring similarity between sentences.
Quiz

• On the next slide, you will see a set of published headlines relating the same scientific study published in 2009. The study suggests that Wolfgang Amadeus Mozart may have died of complications caused by strep throat.

• Take the time to read all these headlines. What interesting (class–related) observations can you make based on your reading?

• What kernels would be most appropriate for clustering all these headlines together? In other words, these kernels should assign high pairwise similarity scores for the headlines in the group.
Did Mozart die of strep throat?

what killed mozart? study suggests strep infection
what killed mozart? strep, study suggests
what killed mozart? study suggest it might have been a strep infection
wstc/wnlk local news what killed mozart? a new theory emerges
strep throat may have led to mozart's death
motor done in by strep throat?
study says mozart died of strep throat
dutch researchers suggest 'super-bug' as cause of mozart's death
new theory on what killed mozart
what killed mozart? study suggests strep
study suggests strep infection killed mozart
'strep throat may have killed mozart'
what killed mozart? study suggests strep
what killed mozart? study hints at complications from strep infection
what killed mozart? study suggests just strep throat
study suggests mozart died of strep infection

mozart may have died from strep throat: study
strep throat theory in mozart's death
medical study suggests mozart died of strep infection died of strep throat?
infected killed mozart?
did strep infection kill mozart?
did a strep infection cause mozart's death?
mozart may have died from strep infection
study reviews mozart's death
infection killed mozart – report
mozart's killer revealed: it was not salieri
infection killed mozart, says study
did poison or strep infection kill mozart?
mozart may have died from strep throat, says study
cause of mozart's death revealed
mozart died from strep
what really killed mozart? possibly strep
mozart may have been killed by strep throat
Answer to Quiz

• Observations
  – All sentences have Mozart
  – The syntax varies a lot (e.g., passive/active voice)
  – They all have some word related to dying (e.g., “kill/die/disease”)

• Suggestions
  – N-gram kernel will probably not work very well here
  – Some semantic information should be encoded in the kernel
  – Possibly, use word2vec
John likes apples, which are green.
Dependencies

- **green**: modifier, child
- **apple**: head, parent
Unionized workers are usually better paid than their non-union counterparts.
The Dependency Tree Kernel
The Syntactic Tree Kernel

Note that siblings are not split up (Collins and Duffy, 2002)
Possible Improvements

• Add semantics (e.g., strong/powerful) [Moschitti et al.]
• Other ideas?
NLP