Recap: how to build such a space

- Solution
  - Low rank matrix approximation

Imagine this is our observed term-document matrix

Imagine this is *true* concept-document matrix

Random noise over the word selection in each document
Recap: Latent Semantic Analysis (LSA)

- Solve LSA by SVD

\[
\hat{Z} = \arg\min_{Z \mid \text{rank}(Z) = k} \|C - Z\|_F \\
= \arg\min_{Z \mid \text{rank}(Z) = k} \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (C_{ij} - Z_{ij})^2} \\
= C_{M \times N}^k
\]

- Procedure of LSA
  1. Perform SVD on document-term adjacency matrix
  2. Construct \( C_{M \times N}^k \) by only keeping the largest \( k \) singular values in \( \Sigma \) non-zero

Map to a lower dimensional space
Introduction to Natural Language Processing

Hongning Wang

CS@UVa
What is NLP?

Arabic text: كلب هو مطاردة صبي في الملعب.

How can a computer make **sense** out of this **string**?

**Morphology**
- What are the basic units of meaning (words)?
- What is the meaning of each word?

**Syntax**
- How are words related with each other?

**Semantics**
- What is the “combined meaning” of words?

**Pragmatics**
- What is the “meta-meaning”? (speech act)

**Discourse**
- Handling a large chunk of text

**Inference**
- Making sense of everything
An example of NLP

A dog is chasing a boy on the playground.

Semantic analysis
Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).

Scared(x) if Chasing(_,x,__).
Scared(b1)

Inference

Lexical analysis (part-of-speech tagging)

Syntactic analysis (Parsing)

A person saying this may be reminding another person to get the dog back...

Pragmatic analysis (speech act)
If we can do this for all the sentences in all languages, then ...

BAD NEWS:
- Unfortunately, we cannot right now.
- General NLP = “Complete AI”
  - Automatically answer our emails
  - Translate languages accurately
  - Help us manage, summarize, and aggregate information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us
NLP is difficult!!!!!!!

• Natural language is designed to make human communication efficient. Therefore,
  – We omit a lot of “common sense” knowledge, which we assume the hearer/reader possesses
  – We keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve

• This makes EVERY step in NLP hard
  – Ambiguity is a “killer”!
  – Common sense reasoning is pre-required
An example of ambiguity

• Get the cat with the gloves.
Examples of challenges

• Word-level ambiguity
  – “design” can be a noun or a verb (Ambiguous POS)
  – “root” has multiple meanings (Ambiguous sense)

• Syntactic ambiguity
  – “natural language processing” (Modification)
  – “A man saw a boy with a telescope.” (PP Attachment)

• Anaphora resolution
  – “John persuaded Bill to buy a TV for himself.” (himself = John or Bill?)

• Presupposition
  – “He has quit smoking.” implies that he smoked before.
Despite all the challenges, research in NLP has also made a lot of progress...
A brief history of NLP

• Early enthusiasm (1950’s): Machine Translation
  – Too ambitious
  – Bar-Hillel report (1960) concluded that fully-automatic high-quality translation could not be accomplished without knowledge (Dictionary + Encyclopedia)

• Less ambitious applications (late 1960’s & early 1970’s): Limited success, failed to scale up
  – Speech recognition
  – Dialogue (Eliza) Shallow understanding
  – Inference and domain knowledge (SHRDLU=“block world”)

• Real world evaluation (late 1970’s – now)
  – Story understanding (late 1970’s & early 1980’s) Knowledge representation
  – Large scale evaluation of speech recognition, text retrieval, information extraction (1980 – now) Robust component techniques
  – Statistical approaches enjoy more success (first in speech recognition & retrieval, later others) Statistical language models

• Current trend:
  – Boundary between statistical and symbolic approaches is disappearing.
  – We need to use all the available knowledge Applications
  – Application-driven NLP research (bioinformatics, Web, Question answering...)
The state of the art

A dog is chasing a boy on the playground

Semantics: some aspects
- Entity/relation extraction
- Word sense disambiguation
- Anaphora resolution

Inference: ???

Speech act analysis: ???

POS Tagging: 97%

Parsing: partial >90%
it's a question, but also an expression of disbelief. Those who get lost driving can use GPS. If you lose your iPhone, there's an app to track it down. Scientists successfully plotted the course for a spacecraft that landed on a speeding asteroid. How did weather affect AirAsia flight? But something goes wrong aboard a 123-foot, 67-ton passenger jet and rescuers must resort to scouring the ocean? "Why is it easier to find an iPhone (than) to find a plane?" one Twitter user, Catalina Buitano, asked. There are dozens of similar questions on social media. They hint at the same sentiment: in a world where people's locations are tracked for everything from map apps to what ads appear on a web browser, why does Big Brother's gaze avoid the skies?

这是一个问题，但也不敢相信的表情。这些谁迷路驾驶可以使用GPS。如果你失去了你的iPhone，有一个应用程序来追查。科学家成功绘制过程中的飞船降落在小行星飞驰。没有天气如何影响亚航的班机吗？但不顺心的事一艘123英尺，67吨重的喷气式客机和救援人员必须求助于大海？“为什么更容易找到一个iPhone（比）找到飞机？”Twitter的用户，卡塔利娜Buitano，问道。有几十个在社交媒体上类似的问题。他们暗示相同的感悟：在这个世界上，人的位置进行跟踪，一切从地图应用程序，以广告出现在网页浏览器，为什么大哥的目光避开天空是什么？
Advances in computation, communication, and data storage techniques in recent decades significantly reduced the cost of data acquisition, leading to an explosion of data generated across different interconnected platforms. Apart from the computational difficulties arise from nonconvex formulations; the sheer volume and spatial disparity of data also pose challenges to traditional learning procedures, which typically require centralized training sets. Reaping the dividend offered by the data deluge, it then urges for the development of collaborative learning methods capable of making inferences from data over the network.

近几十年来，计算、通信和数据存储技术的进步大大降低了数据采集成本，导致跨不同互联平台生成的数据激增。除了计算困难产生于非凸公式;数据的数量和空间差异也对传统的学习程序构成挑战，传统学习程序通常需要集中的培训。通过获得数据泛滥带来的红利，促使开发能够从网络上的数据进行推断的协作学习方法。
Dialog systems

Apple’s siri system

Google search

“Are you human”
Does it matter?
“Yes”
I thought so.
“Answer my question are you human”
Close enough, I’d say.

Google

who is the director of interstellar

Web News Images Videos Shopping More Search tools

About 10,700,000 results (0.52 seconds)

Christopher Nolan
Interstellar, Director
Information extraction
Information extraction

Recently-Learned Facts

<table>
<thead>
<tr>
<th>Instance</th>
<th>Iteration</th>
<th>Date Learned</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>tear_drop_tomatoes is an agricultural product</td>
<td>887</td>
<td>27-nov-2014</td>
<td>93.1</td>
</tr>
<tr>
<td>ryan_mckenzie is a professor</td>
<td>886</td>
<td>21-nov-2014</td>
<td>90.2</td>
</tr>
<tr>
<td>fiorina_161 is a planet</td>
<td>889</td>
<td>07-dec-2014</td>
<td>92.8</td>
</tr>
<tr>
<td>critical_thinking_in_health_science is a cognitive action</td>
<td>886</td>
<td>21-nov-2014</td>
<td>99.0</td>
</tr>
<tr>
<td>fateful_new_year is a monarch</td>
<td>886</td>
<td>21-nov-2014</td>
<td>99.0</td>
</tr>
<tr>
<td>tony_martin has been charged with murder</td>
<td>890</td>
<td>11-dec-2014</td>
<td>100.0</td>
</tr>
<tr>
<td>sen_joe_biden is a U.S. politician who holds the office of vice president</td>
<td>887</td>
<td>27-nov-2014</td>
<td>93.8</td>
</tr>
<tr>
<td>hat is a clothing item to go with blue_jeans</td>
<td>889</td>
<td>07-dec-2014</td>
<td>93.8</td>
</tr>
<tr>
<td>statistics is headquartered in the country the_usa</td>
<td>891</td>
<td>18-dec-2014</td>
<td>98.4</td>
</tr>
<tr>
<td>eoin_colfer wrote the book artemis_fowl</td>
<td>886</td>
<td>21-nov-2014</td>
<td>100.0</td>
</tr>
</tbody>
</table>

YAGO Knowledge Base

CMU Never-Ending Language Learning

CS@UVa CS6501: Text Mining 17
Building a computer that ‘understands’ text: The NLP pipeline
Tokenization/Segmentation

• Split text into words and sentences
  – Task: what is the most likely segmentation/tokenization?

There was an earthquake near D.C. I’ve even felt it in Philadelphia, New York, etc.

There + was + an + earthquake + near + D.C.

I + ve + even + felt + it + in + Philadelphia, + New + York, + etc.
Part-of-Speech tagging

• Marking up a word in a text (corpus) as corresponding to a particular part of speech
  – Task: what is the most likely tag sequence

A + dog + is + chasing + a + boy + on + the + playground

Det Noun Aux Verb Det Noun Prep Det Noun
Named entity recognition

• Determine text mapping to proper names
  – Task: what is the most likely mapping

  Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

  Its initial **Board of Visitors** included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

  **Organization**, **Location**, **Person**
Syntactic parsing

- Grammatical analysis of a given sentence, conforming to the rules of a formal grammar
  - Task: what is the most likely grammatical structure

```
A + dog + is + chasing + a + boy + on + the + playground
```

- Sentence
  - Noun Phrase
  - Complex Verb
  - Noun Phrase
  - Verb Phrase
  - Prep Phrase
  - Noun Phrase
  - Verb Phrase
  - Det
  - Noun
  - Aux
  - Verb
  - Det
  - Noun
  - Prep
  - Det
  - Noun
Relation extraction

• Identify the relationships among named entities
  – Shallow semantic analysis

Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

1. Thomas Jefferson Is_Member_Of Board of Visitors
2. Thomas Jefferson Is_President_Of U.S.
Logic inference

• Convert chunks of text into more formal representations
  – Deep semantic analysis: e.g., first-order logic structures

\[
\exists x \left( \text{Is\_Person}(x) \land \text{Is\_President\_Of}(x, 'U.S.') \land \text{Is\_Member\_Of}(x, 'Board of Visitors') \right)
\]

Its initial **Board of Visitors** included **U.S. Presidents** Thomas Jefferson, James Madison, and James Monroe.
Towards understanding of text

More than a decade ago, Carl Lewis stood on the threshold of what was to become the greatest athletics career in history. He had just broken two of the legendary Jesse Owens' college records, but never believed he would become a corporate icon, the focus of hundreds of millions of dollars in advertising. His sport was still nominally amateur. Eighteen Olympic and World Championship gold medals and 21 world records later, Lewis has become the richest man in the history of track and field -- a multimillionaire.

- Who is Carl Lewis?
- Did Carl Lewis break any records?
Major NLP applications

- Speech recognition: e.g., auto telephone call routing
- Text mining
  - Text clustering
  - Text classification
  - Text summarization
  - Topic modeling
  - Question answering
- Language tutoring
  - Spelling/grammar correction
- Machine translation
  - Cross-language retrieval
  - Restricted natural language
- Natural language user interface
NLP & text mining

• Better NLP => Better text mining

• Bad NLP => Bad text mining?

Robust, shallow NLP tends to be more useful than deep, but fragile NLP.

Errors in NLP can hurt text mining performance…
How much NLP is really needed?

Tasks
- Classification
- Clustering
- Summarization
- Extraction
- Topic modeling
- Translation
- Dialogue
- Question Answering
- Inference
- Speech Act

Dependency on NLP

Scalability
So, what NLP techniques are the most useful for text mining?

- **Statistical NLP** in general.
- The need for high robustness and efficiency implies the dominant use of **simple models**.
What you should know

• Different levels of NLP
• Challenges in NLP
• NLP pipeline