Advanced Multimedia

Intro to NLP & Web Retrieval
Tamara Berg
Course Info

Instructor: Tamara Berg  tlberg@cs.sunysb.edu
Office: 1411 Computer Science
Lectures: Tues/Thurs 3:50-5:10pm, Rm 2205 CS
Office Hours: Tues/Thurs 1:00-2:00pm

Course Webpage:
http://tamaraberg.com/teaching/Spring_12/cse364
Today

• Intro to Natural Language Processing (NLP)
• Meet the web
• Retrieving documents from the web
  Very simple text analysis.
  Use other associated data (links etc) to infer things about documents.
Today

• Intro to Natural Language & Natural Language Processing
• Meet the web
• Retrieving documents from the web
  Very simple text analysis.
  Use other associated data (links etc) to infer things about documents.
The Dream

- It’d be great if machines could
  - Process our email (usefully)
  - Translate languages accurately
  - Help us manage, summarize, and aggregate information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us

- But they can’t:
  - Language is complex, ambiguous, flexible, and subtle
  - Good solutions need linguistics and machine learning knowledge

- So:
What is NLP?

- **Fundamental goal:** *deep* understand of *broad* language
  - Not just string processing or keyword matching!

- **End systems that we want to build:**
  - Ambitious: speech recognition, machine translation, information extraction, dialog interfaces, question answering…
  - Modest: spelling correction, text categorization…
Why study NLP?

- Useful applications...
  - E.g. information retrieval

Topic: Advantages and disadvantages of using potassium hydroxide in any aspect of organic farming, especially...
Why study NLP?

- Useful applications…
  - E.g. question answering systems
    - How many calories are there in a Big Mac?
    - Who is the voice of Miss Piggy?
    - Who was the first American in space?
  - Retrieve not just relevant documents, but return the answer

? \[\rightarrow\] text collection \[\rightarrow\] answer + supporting text

Start System
Why study NLP?

- Useful applications...
  - E.g. summarization

[White et al., 2002]
Original Text

新华网石家庄11月16日电（记者 张涛）11月15日是河北省沧州市的“供暖日”，当地大风、阴雨天，最低气温降至1℃。然而，至少上千户市民家里的暖气仍是冰凉的。原来，这个市今年实施有史以来最大规模的集中供暖“扩面”工程，许多居民小区过去的小锅炉关停、拆除了，而集中供暖却因工程量太大要推迟半个月。

Translated Text

-- Shijiazhuang, November 16 (Xinhua Zhang Tao) November 15 is the city of Cangzhou, Hebei Province "heating Day," local windy, rainy days, the minimum temperature dropped to 1 ℃. However, at least 1,000 members of the public on home heating is still cool. Originally, the city implemented this year's biggest ever focus on heating "expansion of" works, many small residential area in the past a small boiler shutdown, demolition, and the central heating because of too much work should be delayed two weeks.

- SOTA: much better than nothing, but more an understanding aid than a replacement for human translators
- New, better methods e.g. babelfish, translation party
Natural Language

A language that is spoken, signed, or written by humans for general-purpose communication, as distinguished from formal languages (such as computer programming languages or the "languages" used in the study of formal logic) and from constructed languages (esperanto).

Top 10 Languages used on the web

1. English 30.40% 427,436,880
2. Chinese 16.60% 233,216,713
3. Spanish 8.70% 122,349,144
4. Japanese 6.70% 94,000,000
5. French 4.80% 67,315,894
6. German 4.50% 63,611,789
7. Arabic 4.20% 59,810,400
8. Portuguese 4.10% 58,180,960
9. Korean 2.50% 34,820,000
10. Italian 2.40% 33,712,383
11. Rest 15.20% 213,270,757
Natural language on the web

Regular Free text.

Graphics (from Greek γραφικός; see -graphy) are visual presentations on some surface, such as a wall, canvas, computer screen, paper, or stone to brand, inform, illustrate, or entertain. Examples are photographs, drawings, Line Art, graphs, diagrams, typography, numbers, symbols, geometric designs, maps, engineering drawings, or other images. Graphics often combine text, illustration, and color. Graphic design may consist of the deliberate selection, creation, or arrangement of typography alone, as in a brochure, flier, poster, web site, or book without any other element. Clarity or effective communication may be the objective, association with other cultural elements may be sought, or merely, the creation of a distinctive style.

Graphics can be functional or artistic. The latter can be a recorded version, such as a photograph, or an interpretation by a scientist to highlight essential features, or an artist, in which case the distinction with imaginary graphics may become blurred.
Doctor Nikola shows a fork that was removed from an Israeli woman who swallowed it while trying to catch a bug that flew in to her mouth, in Poriah Hospital northern Israel July 10, 2003. Doctors performed emergency surgery and removed the fork. (Reuters)
Natural language on the web

Captions – natural language, but highly stylized & directly associated with pictures.

Doctor Nikola shows a fork that was removed from an Israeli woman who swallowed it while trying to catch a bug that flew in to her mouth, in Poriah Hospital northern Israel July 10, 2003. Doctors performed emergency surgery and removed the fork. (Reuters)
Natural language on the web

Speech - with the explosion of video on the web the amount of speech is also growing quickly.
Natural language on the web

Tags/keywords

– Folksonomy is the practice and method of collaboratively creating and managing tags to annotate and categorize content.

– Usually, freely chosen keywords are used instead of a controlled vocabulary.

– Became popular on the Web around 2004 as part of social software applications including social bookmarking and annotating photographs. Tagging allows non-expert users to collectively classify and find information.
Web Pages

• Contain all of these kinds of language + some additional constructed items like hyperlinks, title, url etc.

• For now we will focus on standard written natural language + special web related items.

• Later in the class we will look at tags, speech etc.
Today

• Intro to Natural Language & Natural Language Processing
• Meet the web
• Retrieving documents from the web
  Very simple text analysis.
  Use other associated data (links etc) to infer things about documents.
How big is the web?

- The first Google index in 1998 already had 26 million pages.
- By 2000 the Google index reached the one billion mark.
- July 25, 2008 – Google announced that search had discovered one trillion unique URLs.
How many people use the web?

Population Percentage online (march 31, 2011):
- Africa – 11.4%
- Asia – 23.8%
- Europe – 58.3%
- Middle East – 31.7%
- North America – 78.3%
- Latin America – 36.2%
- Oceania/Australia – 60.1%

Internetworldstats.com

Google searches per day (comscore):
- 1998 10,000
- 1999 500,000
- 2008 235 million
- 2010 3 billion
What is the shape of the web?

Medusa Model of the web at the autonomous system level

“A model of Internet topology using k-shell decomposition,” Shai Carmi et al PNAS
What is the shape of the web?

Increasing number of connections
What is the shape of the web?
The Web as a Directed Graph

Assumption 1: A hyperlink between pages denotes author perceived relevance (quality signal)

Assumption 2: The anchor of the hyperlink describes the target page (textual context)
How hard is it to go from one page to another?

• Over 75% of the time there is no directed path from one random web page to another.
How hard is it to go from one page to another?

• Over 75% of the time there is no directed path from one random web page to another.
• When a directed path exists its average length is 16 clicks.
• When an undirected path exists its average length is 7 clicks.
How hard is it to go from one page to another?

• Over 75% of the time there is no directed path from one random web page to another.
• When a directed path exists its average length is 16 clicks.
• When an undirected path exists its average length is 7 clicks.
• Short average path between pairs of nodes is characteristic of a small-world network (“six degrees of separation” Stanley Milgram).
Today

• Intro to Natural Language & Natural Language Processing
• Meet the web
• Retrieving documents from the web
  Very simple text analysis.
  Use other associated data (links etc) to infer things about documents.
Information Retrieval

- Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web.
Web search vs. Info Retrieval

- The scale of web search is way beyond traditional information retrieval.
- The web is very dynamic.
- The web contains an enormous amount of duplication.
- The quality of web pages is not uniform.
- The range of topics on the web is open.
- The web is globally distributed.
- Users typical habits are different (short queries, inspect only top-10 pages).
- The web is hypertextual.
Taxonomy of Web Queries

1. Navigational. The immediate intent is to reach a particular site.

2. Informational. The intent is to acquire some information assumed to be present on one or more web pages.

3. Transactional. The intent is to perform some web-mediated activity.
Navigational Queries

The purpose of such queries is to reach a particular site that the user has in mind, either because they visited in the past or because they assume that such a site exists.

Examples:

• Greyhound Bus. Probable target http://www.greyhound.com
• National car rental. Probable target http://www.nationalcar.com
• American airlines home. Probable target http://www.aa.com
• Don Knuth. Probable target http://www-cs-faculty.stanford.edu/~knuth/

Usually only have one “right” answer
Informational Queries

Want to find information about a particular topic.

What is different on the web is that many informational queries are extremely wide, for instance cars or San Francisco, while some are narrow, for instance normocytic anemia, or Scoville heat units.

It is interesting to note, that in almost 15% of all searches the desired target is a good collection of links on the subject, rather than a good document.
Transactional Queries

• The purpose of such queries is to reach a site where further interaction will happen. This interaction constitutes the transaction defining these queries.
• The main categories for such queries are shopping, finding various web-mediated services, downloading various type of file (images, songs, etc), accessing certain data-bases (e.g. Yellow Pages type data), finding servers (e.g. for gaming) etc.
Query Breakdown

- Navigational 20%
- Informational 48%
- Transactional 30%

Estimated using query log analysis of user behavior on 400 queries extracted from the daily AltaVista query log.
The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page
“The ultimate search engine would understand exactly what you mean and give back exactly what you want.” - Larry Page

Google – misspelling of googol $= 10^{100}$
The Google Search Engine

Founded 1998 (1996) by two Stanford students

Originally academic / research project that later became a commercial tool

Distinguishing features (then!?):
- Special (and better) ranking
- Speed
- Size
The web in 1997

Internet was growing very quickly

• “Junk results often wash out any results that a user is interested in. In fact, as of November 1997, only one of the top four commercial search engines finds itself (returns its own search page in response to its name in the top ten results).”
The web in 1997

Internet was growing very quickly

• “Junk results often wash out any results that a user is interested in. In fact, as of November 1997, only one of the top four commercial search engines finds itself (returns its own search page in response to its name in the top ten results).”

Need high precision in the top results
Google’s first search engine
Components of Web Search Service

Components
• Web crawler
• Indexing system
• Search system
• Advertising system

Considerations
• Economics
• Scalability
• Legal issues
Web Searching: Architecture

- Documents stored on many Web servers are indexed in a single central index.
- The central index is implemented as a single system on a very large number of computers

Examples: Google, Yahoo!

Slide from William Y. Arms
What is a Web Crawler?

**Web Crawler**

- A program for downloading web pages.
- Given an initial set of seed URLs, it recursively downloads every page that is linked from pages in the set.
- A **focused** web crawler downloads only those pages whose content satisfies some criterion.

*Also known as a web spider*
Simple Web Crawler Algorithm

**Basic Algorithm**

Let $S$ be set of URLs to pages waiting to be indexed. Initially $S$ is a set of known seeds.

Take an element $u$ of $S$ and **retrieve** the page, $p$, that it references.

Parse the page $p$ and **extract** the set of URLs $L$ it has links to.

**Update** $S = S + L - u$

**Repeat** as many times as necessary.

*[Large production crawlers may run continuously]*
Indexing the Web Goals: Precision

*Short queries applied to very large numbers of items leads to large numbers of hits.*

- Goal is that the first 10-100 hits presented should satisfy the user's information need
  -- requires ranking hits in order that fits user's requirements
- Recall is not an important criterion

*Completeness of index is not an important factor.*

- Comprehensive crawling is unnecessary
Concept of Relevance and Importance

Document measures

**Relevance**, as conventionally defined, is binary (relevant or not relevant). It is usually estimated by the similarity between the terms in the query and each document.

**Importance** measures documents by their likelihood of being useful to a variety of users. It is usually estimated by some measure of popularity.

Web search engines rank documents by a weighted combination of estimates of relevance and importance.
Relevance

• Words in document (stored in inverted index)
• Location information – for use of proximity in multi-word search.
• In page title, page url?
• Visual presentation details – font size of words, words in bold.
The source of Document A contains the marked-up text:

```
<a href="http://www.cis.cornell.edu/">The Faculty of Computing and Information Science</a>
```

The **anchor text**:

The Faculty of Computing and Information Science can be considered **descriptive metadata** about the document:

http://www.cis.cornell.edu/
Importance - PageRank Algorithm

Used to estimate popularity of documents

Concept:

The rank of a web page is higher if many pages link to it.

Links from highly ranked pages are given greater weight than links from less highly ranked pages.
IR score

• Weights for each relevance feature type
  \[ w = [w_1 \ w_2 \ w_3 \ w_4 \ ...] \]

• Relevance score for each feature type
  \[ r = [r_1 \ r_2 \ r_3 \ r_4 \ ...] \]

• IR Score is weighted combination of relevance score
  \[ \text{IR} = \text{dot}(w, r) \]

• Final rank computed from IR score and PageRank
Importance - PageRank Algorithm

**Used to estimate popularity of documents**

**Concept:**

The rank of a web page is higher if many pages link to it.

Links from highly ranked pages are given greater weight than links from less highly ranked pages.
Intuitive Model (Basic Concept)

Basic (no damping)
A user:
1. Starts at a random page on the web
2. Selects a random hyperlink from the current page and jumps to the corresponding page
3. Repeats Step 2 a very large number of times

Pages are ranked according to the relative frequency with which they are visited.
PageRank
Basic Algorithm: Matrix Representation

<table>
<thead>
<tr>
<th>Citing page (from)</th>
<th>P_1</th>
<th>P_2</th>
<th>P_3</th>
<th>P_4</th>
<th>P_5</th>
<th>P_6</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>P_2</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>P_3</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>P_4</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>P_5</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>P_6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Cited page (to)

Number

4  2  1  1  3  1  1

Slide from William Y. Arms
Basic Algorithm: Normalize by Number of Links from Page

<table>
<thead>
<tr>
<th>Citing page</th>
<th>P₁</th>
<th>P₂</th>
<th>P₃</th>
<th>P₄</th>
<th>P₅</th>
<th>P₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>P₁</td>
<td></td>
<td></td>
<td></td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₂</td>
<td>0.25</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₃</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₄</td>
<td>0.25</td>
<td>0.5</td>
<td></td>
<td>0.33</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P₅</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₆</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.33</td>
<td></td>
</tr>
</tbody>
</table>

| Number | 4 | 2 | 1 | 1 | 3 | 1 |

Slide from William Y. Arms
Basic Algorithm: Normalize by Number of Links from Page

<table>
<thead>
<tr>
<th>Cited page</th>
<th>P₁</th>
<th>P₂</th>
<th>P₃</th>
<th>P₄</th>
<th>P₅</th>
<th>P₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>P₁</td>
<td></td>
<td></td>
<td></td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₂</td>
<td>0.25</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₃</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₄</td>
<td>0.25</td>
<td>0.5</td>
<td>0.33</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P₅</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.33</td>
</tr>
<tr>
<td>P₆</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number: 4 2 1 1 3 1

Normalized link matrix = B
Basic Algorithm: Weighting of Pages

Initially all pages have weight $1/n$

$$w_0 = \begin{bmatrix} 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \end{bmatrix}$$
Basic Algorithm: Weighting of Pages

Initially all pages have weight $1/n$

Recalculate weights

$$ w_0 = \begin{bmatrix} 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \end{bmatrix} $$

$$ w_1 = Bw_0 = \begin{bmatrix} 0.06 \\ 0.21 \\ 0.29 \\ 0.35 \\ 0.04 \\ 0.06 \end{bmatrix} $$

If the user starts at a random page, the $j$th element of $w_1$ is the probability of reaching page $j$ after one step.
Basic Algorithm: Weighting of Pages

Initially all pages have weight $1/n$

Recalculate weights

$$w_0 = \begin{bmatrix} 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \\ 0.17 \end{bmatrix}$$

$$w_1 = Bw_0 = \begin{bmatrix} 0.06 \\ 0.21 \\ 0.29 \\ 0.35 \\ 0.04 \\ 0.06 \end{bmatrix}$$

If the user starts at a random page, the $j^{th}$ element of $w_1$ is the probability of reaching page $j$ after one step.

Slide from William Y. Arms
Basic Algorithm: Iterate

Iterate: \( w_k = B w_{k-1} \)

<table>
<thead>
<tr>
<th>( w_0 )</th>
<th>( w_1 )</th>
<th>( w_2 )</th>
<th>( w_3 )</th>
<th>... converges to ...</th>
<th>( w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>0.06</td>
<td>0.01</td>
<td>0.01</td>
<td>-&gt;</td>
<td>0.00</td>
</tr>
<tr>
<td>0.17</td>
<td>0.21</td>
<td>0.32</td>
<td>0.47</td>
<td>-&gt;</td>
<td>0.40</td>
</tr>
<tr>
<td>0.17</td>
<td>0.29</td>
<td>0.46</td>
<td>0.34</td>
<td>-&gt;</td>
<td>0.40</td>
</tr>
<tr>
<td>0.17</td>
<td>0.35</td>
<td>0.19</td>
<td>0.17</td>
<td>-&gt;</td>
<td>0.20</td>
</tr>
<tr>
<td>0.17</td>
<td>0.04</td>
<td>0.01</td>
<td>0.00</td>
<td>-&gt;</td>
<td>0.00</td>
</tr>
<tr>
<td>0.17</td>
<td>0.06</td>
<td>0.01</td>
<td>0.00</td>
<td>-&gt;</td>
<td>0.00</td>
</tr>
</tbody>
</table>

At each iteration, the sum of the weights is 1.
Special Cases of Hyperlinks on the Web

There is no link out of \{2, 3, 4\}
Google PageRank with Damping

A user:

1. Starts at a random page on the web
2a. With probability \(1-d\), selects any random page and jumps to it
2b. With probability \(d\), selects a random hyperlink from the current page and jumps to the corresponding page
3. Repeats Step 2a and 2b a very large number of times

Pages are ranked according to the relative frequency with which they are visited.

[For dangling nodes, always follow 2a.]
The PageRank Iteration

The **basic method** iterates using the **normalized link matrix**, $B$.

$$w_k = Bw_{k-1}$$

This $w$ is an eigenvector of $B$

**PageRank** iterates using a damping factor. The method iterates:

$$w_k = (1 - d)w_0 + dBw_{k-1}$$

$w_0$ is a vector with every element equal to $1/n$. 
The PageRank Iteration

The iteration expression with damping can be re-written.

Let \( R \) be a matrix with every element equal to \( 1/n \)

\[
Rw_{k-1} = w_0 \quad \text{(The sum of the elements of } w_{k-1} \text{ equals 1)}
\]

Let \( G = dB + (1-d)R \) (\( G \) is called the Google matrix)

The iteration formula

\[
w_k = (1-d)w_0 + dBw_{k-1}
\]

is equivalent to

\[
w_k = Gw_{k-1}
\]

so that \( w \) is an eigenvector of \( G \)
Iterate with Damping

Iterate: $w_k = Gw_{k-1}$ ($d = 0.7$)

\[
\begin{bmatrix}
0.17 \\
0.17 \\
0.17 \\
0.17 \\
0.17 \\
0.17 \\
0.17
\end{bmatrix}
\begin{bmatrix}
0.09 \\
0.20 \\
0.26 \\
0.30 \\
0.08 \\
0.07 \\
0.09
\end{bmatrix}
\begin{bmatrix}
0.07 \\
0.24 \\
0.34 \\
0.30 \\
0.22 \\
0.07 \\
0.07
\end{bmatrix}
\begin{bmatrix}
0.07 \\
0.30 \\
0.30 \\
0.21 \\
0.06 \\
0.06 \\
0.06
\end{bmatrix}
\rightarrow
\begin{bmatrix}
0.06 \\
0.28 \\
0.31 \\
0.22 \\
0.06 \\
0.06 \\
0.06
\end{bmatrix}
\]
Choice of $d$

Conceptually, values of $d$ that are close to 1 are desirable as they emphasize the link structure of the Web graph, but...

- The rate of convergence of the iteration decreases as $d$ approaches 1.
- The sensitivity of PageRank to small variations in data increases as $d$ approaches 1.

It is reported that Google uses a value of $d = 0.85$ and that the computation converges in about 50 iterations.