

Text Retrieval

- Introduction
- Text Retrieval Tasks
- Evaluation
- The Vector Space Model
- Advanced Techniques



1

CIS-530

Readings

- Manning Chapter: Text Retrieval (Selections)
- Vorhees & Harman (Bulkpack)



2

CIS-530

Text Retrieval: Overview

- **Goal:**
 - Use a collection of documents to generate a response to a query.
- **Prototypical example:**
 - Web search engine



3

CIS-530

Text Retrieval Tasks

- **Ad-hoc retrieval**
 - Given: Large fixed document set,
 - Goal: Find documents that answer queries.
- **Filtering**
 - Given: Fixed query
 - Goal: Decide whether documents answer that query.
- **Question-Answering**
 - Given: Large fixed document set
 - Goal: Find short passages (~ 1 sentence) that answer questions.



4

CIS-530

Text Retrieval Tasks: Classification

There are many different kinds of text retrieval.

- **Variables for Classification:**

- Is the query known ahead of time?
- Is the corpus known ahead of time?
- What kind of query is used?
- What kind of corpus is used?
- What kind of response should be generated?



Text Retrieval Tasks: Classification (Cont'd)

- **Example classifications:**

Task	Known ahead of time?	Corpus	Query	Response
Ad-Hoc	Corpus		Short – Medium	Document
Question Answering	Corpus		Short	Sentence
Filtering	Query		Long	Document
Cross-Language	Corpus	Multilingual	Short	Document
Speech	Corpus	Speech	Short – Medium	Document



TREC

- **Text REtrieval Conference**

- Annual workshop to foster research in text retrieval
- 50+ groups compete for high performance.

- **16 Tracks = types of text retrieval**

- **7 tracks were active in TREC-9**
- | | | |
|-------------|------------------|-------------------|
| Ad-Hoc | Database Merging | High Precision |
| Routing | Filtering | Very Large Corpus |
| Interactive | Chinese | Query |
| Spanish | NLP | Cross-Language |
| Confusion | Speech | Web |



Evaluation

- **Precision/Recall: Review**

	Target	¬Target
Selected	True positive	False positive
¬Selected	False negative	True negative

- **Precision** = $\frac{tp}{(tp+fp)}$

- What proportion of selected items are correct?

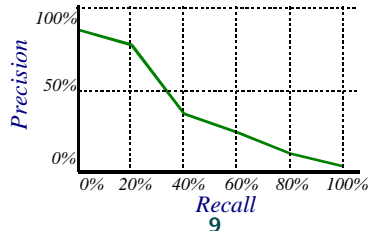
- **Recall** = $\frac{tp}{(tp+fn)}$

- What proportion of target items are selected?



Evaluation: Precision/Recall Graphs

- Text retrieval generates multiple responses
- Consider the first n responses for various n
 - $n=0$: precision=100%; recall=0%
 - $n=N$: precision=0%; recall=100%
- Graph precision vs. recall at various *cutoffs*



CIS-530

9

Evaluation: Summary Measures

- Uninterpolated average precision
 - P = positions at which we got a true positive
 - average precision = $\text{Avg}(\text{precision at cutoff } p)$
 $p \in P$
- F Measure
 - Weighted harmonic mean of precision and recall.

$$\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}$$



10

CIS-530

Evaluation: Test Set Construction

- Which documents are relevant?
 - Need human evaluation
 - New evaluation for each query
- Need a *complete set* of relevant documents
 - Can't find recall without a complete set
- But there are too many documents to evaluate!



CIS-530

11

Evaluation: Pooling

- TREC: Construct a test set that is "approximately correct"
 - Assume that all relevant documents are returned by at least one text retrieval system
- Manually evaluate the top 100 documents returned by each system.
- No bias against any competing system.
- Small bias against systems that did not compete.



12

CIS-530

The Vector Space Model

Represent each document as a sparse vector \vec{x}

- Each dimension corresponds to a single vocabulary item:
 - 1 if the vocabulary item is in the document
 - 0 if the vocabulary item is not in the document

$d = \text{document (multiset of words)}$ $\vec{x} = \langle x_1, x_2, \dots, x_n \rangle$

$V = w_1, w_2, \dots, w_n = \text{vocabulary}$

$w_i = \text{word}$

$$x_i = \begin{cases} 1 & \text{if } w_i \in d \\ 0 & \text{if } w_i \notin d \end{cases}$$

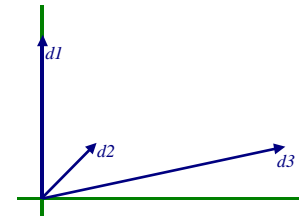


13

CIS-530

Comparing Documents

- Two documents are similar if they have similar term vectors.

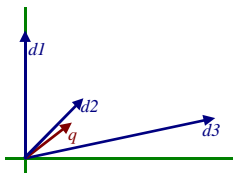


14

CIS-530

Comparing Documents & Queries

- Treat the query as a very small document.
- Construct a vector representation of the query.



- Query vectors are typically shorter than document vectors.



15

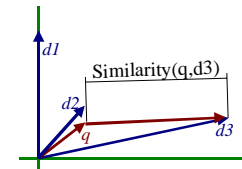
CIS-530

Comparing Term Vectors

What makes term vectors similar?

- **Attempt 1:** Term vectors are similar if their difference is small.

$$\text{similarity}_1(\vec{x}_1, \vec{x}_2) = |\vec{x}_1 - \vec{x}_2|$$



- Does document length matter?
- We only care about the *relative frequency* of each term.



16

CIS-530

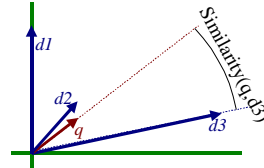
Comparing Term Vectors Cont'd: The Cosine Measure

What makes term vectors similar?

- **Attempt 2:** Term vectors are similar if the angle between them is small.

$$\text{similarity}_2(\vec{x}_1, \vec{x}_2) = \cos(\vec{x}_1, \vec{x}_2)$$

$$\cos(\vec{x}_1, \vec{x}_2) = \frac{\vec{x}_1 \cdot \vec{x}_2}{|\vec{x}_1| \times |\vec{x}_2|}$$



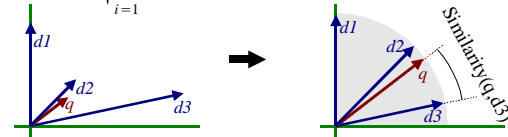
17

CIS-530

Normalizing Term Vectors

- Cosines are easier to compute if we first *normalize* all document vectors.

$$\vec{x}' = \frac{\vec{x}}{|\vec{x}|} = \frac{\vec{x}}{\sqrt{\sum_{i=1}^n x_i^2}}$$



$$\text{similarity}_2(\vec{x}_1', \vec{x}_2') = \vec{x}_1' \cdot \vec{x}_2'$$



18

CIS-530

Term Weighting

- Some terms are more informative than others.
- Can we get a better similarity measure if we use different vectors?

$$\vec{x} = \langle x_1, x_2, \dots, x_n \rangle$$

- Change our definition of x_i

$$x_i = \begin{cases} 1 & \text{if } w_i \in d \\ 0 & \text{if } w_i \notin d \end{cases}$$

- Increase x_i for more informative terms.



19

CIS-530

Term Weighting (Continued) Term Frequency

- **Term frequency:** $tf_i = \text{number of occurrences of } w_i \in d$
- Higher term frequency \rightarrow term is more relevant to the document.
- Weight a term proportionally to its term frequency? $x_i = tf_i$
- But a word appearing 3 times as often is not 3 times as relevant: use smoothing.

$$x_i = \begin{cases} 1 + \log(tf_i) & \text{if } w_i \in d \\ 0 & \text{if } w_i \notin d \end{cases} \quad \text{or} \quad x_i = \sqrt{tf_i}$$



20

CIS-530

Term Weighting (Continued) Inverse Document Frequency

- Document Frequency:

$df_i = \text{number of documents containing } w_i$

- Higher document frequency \rightarrow term is less informative.
- Weight a term inversely to its document frequency.

$$x_i = \begin{cases} (1 + \log(tf_i)) \log\left(\frac{N}{df_i}\right) & \text{if } w_i \in d \\ 0 & \text{if } w_i \notin d \end{cases}$$

- (use smoothing)



21

CIS-530

Latent Semantic Indexing

- Vector model assumes that term contributions to document similarity are *independent*.
- Many terms are not independent.
 - "star chart" is similar to "astral map"
- Project document vectors into a new vector space.
 - In the new vector space, terms that are semantically similar are close to each other.
 - "Latent" semantic dimensions



22

CIS-530

Using NLP Techniques

- Traditional wisdom: linguistic knowledge does not improve text retrieval performance.
 - Especially true for ad-hoc retrieval.
- But:
 - Linguistic knowledge useful for other tasks (e.g., question answering, cross-language text retrieval)
 - Clearly a *temporary* fact: it is impossible to get 100% performance without linguistic knowledge.



23

CIS-530

Using NLP Techniques: Words

- Add new terms:
 - Find terms with collocation analysis
 - Synonym expansion: WordNet
- Make terms more precise:
 - Word sense disambiguation
- Merge terms with the same meaning:
 - Stemming
- Clean up a "messy" corpus:
 - Language modelling
 - Spell-check



24

CIS-530

Using NLP Techniques: Syntax

- **Find new terms:**
 - **Entity extraction**
 - **Relationship extraction**
- **Narrow the search space:**
 - **Question parsing**
 - Determine what type of answer we expect
 - **Entity tagging**
 - Tag entities with types (e.g., company, person, date)
 - **Relationship extraction**
 - Search for partial relationships (e.g., "John likes ?x")

