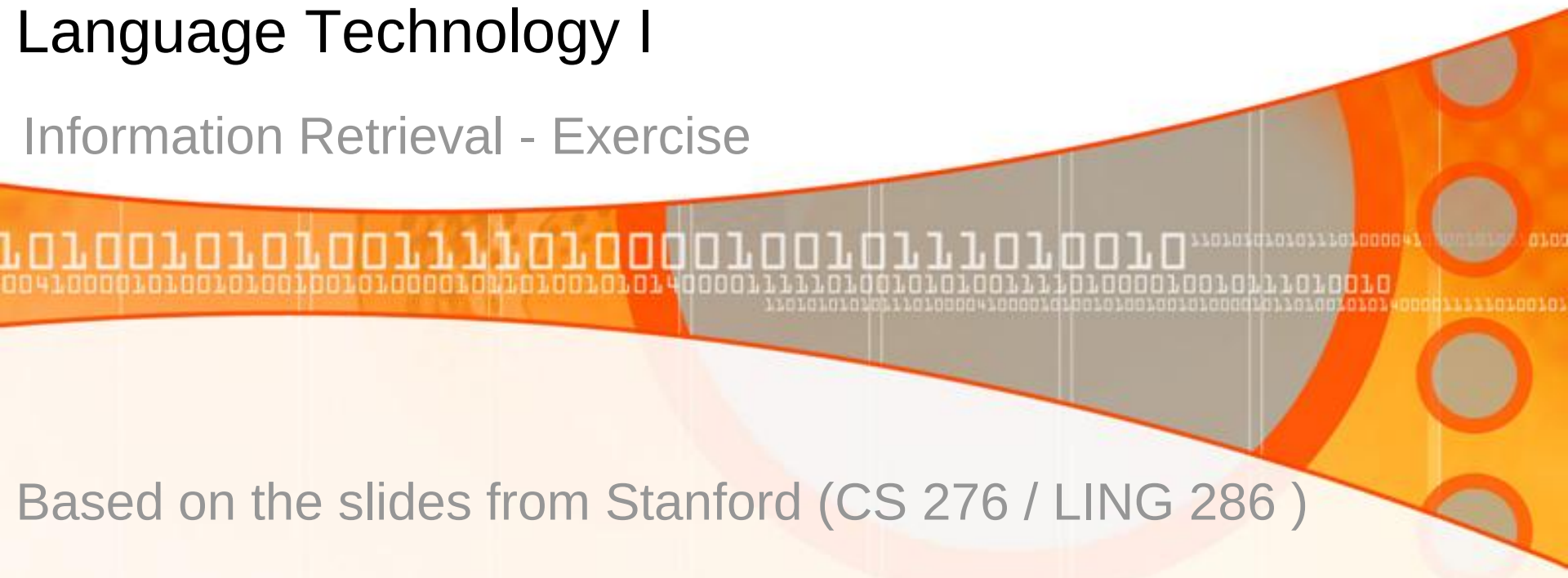
Language Technology I

Information Retrieval - Exercise



Based on the slides from Stanford (CS 276 / LING 286)

Recap: Unranked retrieval evaluation: Precision and Recall

- **Precision**: fraction of retrieved docs that are relevant = P(relevant|retrieved)
- **Recall**: fraction of relevant docs that are retrieved = P(retrieved|relevant)

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- *Precision P = tp/(tp + fp)*
- *Recall R = tp/(tp + fn)*

Recap: A combined measure: *F*

- Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- People usually use balanced F_1 measure
  - *i.e., with* $\beta = 1$

Evaluation at large search engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the web
- Search engines often use precision at top k, e.g., k = 10
- . . . or measures that reward you more for getting rank 1 right than for getting rank 10 right.
  - *NDCG (Normalized Cumulative Discounted Gain)*

Overview

- Improving results
  - *For high recall. E.g., searching for* aircraft *doesn't match with* plane; *nor* thermodynamic *with* heat
- Options for improving results…
  - *Focus on relevance feedback*
  - *The complete landscape*
    - *Global methods*
      - *Query expansion*
        - *Thesauri*
        - *Automatic thesaurus generation*
    - *Local methods*
      - *Relevance feedback*
      - *Pseudo relevance feedback*

Relevance Feedback

- Relevance feedback: user feedback on relevance of docs in initial set of results
  - *User issues a (short, simple) query*
  - *The user marks some results as relevant or non-relevant.*
  - *The system computes a better representation of the information need based on feedback.*
  - *Relevance feedback can go through one or more iterations.*
- Idea: it may be difficult to formulate a good query when you don't know the collection well, so iterate

Initial query/results

- Initial query: *New space satellite applications*
  + 1. 0.539, 08/13/91, *NASA Hasn't Scrapped Imaging Spectrometer*
  + 2. 0.533, 07/09/91, *NASA Scratches Environment Gear From Satellite Plan*
  3. 0.528, 04/04/90, *Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes*
  4. 0.526, 09/09/91, *A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget*
  5. 0.525, 07/24/90, *Scientist Who Exposed Global Warming Proposes Satellites for Climate Research*
  6. 0.524, 08/22/90, *Report Provides Support for the Critics Of Using Big Satellites to Study Climate*
  7. 0.516, 04/13/87, *Arianespace Receives Satellite Launch Pact From Telesat Canada*
  + 8. 0.509, 12/02/87, *Telecommunications Tale of Two Companies*
- User then marks relevant documents with "+".

Expanded query after relevance feedback

- 2.074 new
- 30.816 satellite
- 5.991 nasa
- 4.196 launch
- 3.516 instrument
- 3.004 bundespost
- 2.790 rocket
- 2.003 broadcast
- 0.836 oil
- 15.106 space
- 5.660 application
- 5.196 eos
- 3.972 aster
- 3.446 arianespace
- 2.806 ss
- 2.053 scientist
- 1.172 earth
- 0.646 measure

Results for expanded query

1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan 2
2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer 1
3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit
5. 0.492, 12/02/87, Telecommunications Tale of Two Companies 8
6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use
7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

Relevance Feedback: Assumptions

- User has sufficient knowledge for initial query.
- Relevance prototypes are “well-behaved”.
  - *Term distribution in relevant documents will be similar*
  - *Term distribution in non-relevant documents will be different from those in relevant documents*
    - *All relevant documents are tightly clustered around a single prototype.*
    - *Similarities between relevant and irrelevant documents are small*

Relevance Feedback: Problems

- Long queries are inefficient for typical IR engine.
  - *Long response times for user.*
  - *High cost for retrieval system.*
  - *Partial solution:*
    - *Only reweight certain prominent terms*
      - *Perhaps top 20 by term frequency*
- Users are often reluctant to provide explicit feedback
- It's often harder to understand why a particular document was retrieved after applying relevance feedback

Relevance Feedback on the Web

- Some search engines offer a similar/related pages feature (this is a trivial form of relevance feedback)
  - *Google (link-based)*
  - *Altavista*
  - *Stanford WebBase*
- But some don't because it's hard to explain to average user:
  - *Alltheweb*
  - *msn live.com*
  - *Yahoo*
- Excite initially had true relevance feedback, but abandoned it due to lack of use.

Pseudo relevance feedback

- Pseudo-relevance feedback automates the “manual” part of true relevance feedback.
- Pseudo-relevance algorithm:
  - *Retrieve a ranked list of hits for the user’s query*
  - *Assume that the top k documents are relevant.*
  - *Do relevance feedback (e.g., Rocchio)*
- Works very well on average
- But can go horribly wrong for some queries.
- Several iterations can cause query drift.

Query Expansion

- In relevance feedback, users give additional input (relevant/non-relevant) on documents, which is used to reweight terms in the documents
- In query expansion, users give additional input (good/bad search term) on words or phrases

How do we augment the user query?

- Manual thesaurus
  - *E.g. MedLine: physician, syn: doc, doctor, MD, medico*
  - *Can be query rather than just synonyms*
- Global Analysis: (static; of all documents in collection)
  - *Automatically derived thesaurus*
    - *(co-occurrence statistics)*
  - *Refinements based on query log mining*
    - *Common on the web*
- Local Analysis: (dynamic)
  - *Analysis of documents in result set*

Thesaurus-based query expansion

- For each term, *t*, in a query, expand the query with synonyms and related words of *t* from the thesaurus
  - *feline → feline cat*
- May weight added terms less than original query terms.
- Generally increases recall
- Widely used in many science/engineering fields
- May significantly decrease precision, particularly with ambiguous terms.
  - *"interest rate" → "interest rate fascinate evaluate"*
- There is a high cost of manually producing a thesaurus
  - *And for updating it for scientific changes*

Automatic Thesaurus Generation

- Attempt to generate a thesaurus automatically by analyzing the collection of documents
- Fundamental notion: similarity between two words
- Definition 1: Two words are similar if they co-occur with similar words.
- Definition 2: Two words are similar if they occur in a given grammatical relation with the same words.
- You can harvest, peel, eat, prepare, etc. apples and pears, so apples and pears must be similar.
- Co-occurrence based is more robust, grammatical relations are more accurate.

Automatic Thesaurus Generation Example

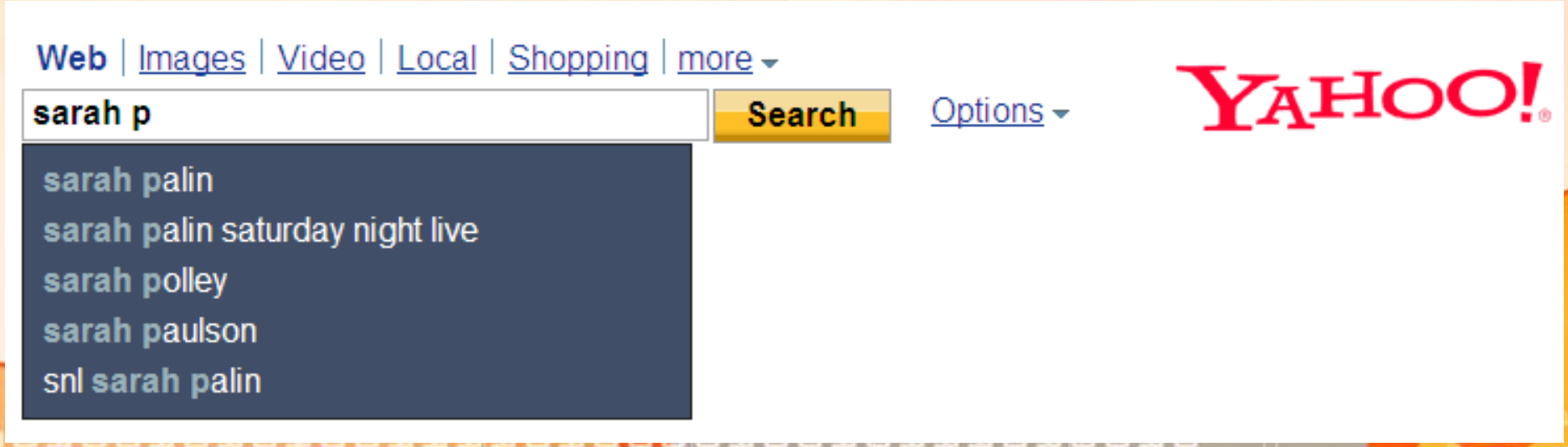
word	ten nearest neighbors
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed slig
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs gazec
Makeup	repellent lotion glossy sunscreen Skin gel p
mediating	reconciliation negotiate cease conciliation p
keeping	hoping bring wiping could some would othe
lithographs	drawings Picasso Dali sculptures Gauguin
pathogens	toxins bacteria organisms bacterial parasite
senses	grasp psyche truly clumsy naive innate awl

Automatic Thesaurus Generation Discussion

- Quality of associations is usually a problem.
- Term ambiguity may introduce irrelevant statistically correlated terms.
  - "Apple computer" □ "Apple red fruit computer"
- Problems:
  - False positives: Words deemed similar that are not
  - False negatives: Words deemed dissimilar that are similar
- Since terms are highly correlated anyway, expansion may not retrieve many additional documents.

Query assist

- Generally done by query log mining
- Recommend frequent recent queries that contain partial string typed by user
- A ranking problem! View each prior query as a doc – Rank-order those matching partial string



Indirect relevance feedback

- On the web, DirectHit introduced a form of indirect relevance feedback.
- DirectHit ranked documents higher that users look at more often.
  - *Clicked on links are assumed likely to be relevant*
    - *Assuming the displayed summaries are good, etc.*
- Globally: Not necessarily user or query specific.
  - *This is the general area of clickstream mining*
- Today – handled as part of machine-learned ranking