LING 120, Fall 2017
Language and Computers

Instructor: Sowmya Vajjala

Iowa State University, USA

25 September 2017

Class outline

1. Assignment 2 discussion
2. Recap of Week 5
3. Searching in large collections of text corpora (not the web)
4. Introduction to Regular Expressions
5. Assignment 3 description

Assignment 2 discussion

- Q1: isolated errors
  - Some errors have an explanation (learing - could be a missing letter)
  - Some errors are random, and could have occurred just because it is a test scenario and students were typing in a hurry
  - Tools offer various suggestions: learning, leering, leaning, clearing etc.

Assignment 2 discussion

- Q1: isolated errors
  - Some errors have an explanation (learing - could be a missing letter)
  - Some errors are random, and could have occurred just because it is a test scenario and students were typing in a hurry
  - Tools offer various suggestions: learning, leering, leaning, clearing etc.
- Q2: contexual errors
  - Google seems to offer better suggestions for contexual errors (except "sap opera", why?).
  - Grammarly seems to be good at that too (except "golf war", what could be the reason?
  - "I went their house" - strangely, none got them right.
  - MS Word does not seem to do any contextual error correction

Quick recap of last week

Topics discussed

- Searching through structured data
- Searching through unstructured data: searching the web
- Language issues in search
- Indexing the web: term document matrix and inverted index
- Ranking the web: Page Rank and other features

A few questions

- What are ”stop words”?

A few questions

- What are ”stop words”?
- Why is stop word removal done in search?

A few questions

- What are "stop words"?
- Why is stop word removal done in search?
- Let T be the total number of pages retrieved for a given search query, R be the number of relevant results among these, and A be the number of actual relevant results on the web.
- What is the difference between T and A?

A few questions

- What are "stop words"?
- Why is stop word removal done in search?
- Let T be the total number of pages retrieved for a given search query, R be the number of relevant results among these, and A be the number of actual relevant results on the web.
- What is the difference between T and A?
- What is R/A? What is R/T?

A few questions

- What are "stop words"?
- Why is stop word removal done in search?
- Let T be the total number of pages retrieved for a given search query, R be the number of relevant results among these, and A be the number of actual relevant results on the web.
- What is the difference between T and A?
- What is R/A? What is R/T?
- What is desirable? High precision or High recall?

Last class' Exercise

- Work in groups of 3 and submit a solution to the problem in the handout.
- You can also submit this online on Canvas.
- question url: `http://nacloweb.org/resources/problems/2007/N2007-B.pdf`

Last class' Exercise

- Work in groups of 3 and submit a solution to the problem in the handout.
- You can also submit this online on Canvas.
- question url: `http://nacloweb.org/resources/problems/2007/N2007-B.pdf`
- solution url: `http://www.education.rec.ri.cmu.edu/fire/naclo/pdfs/pooh-encyclopedia-solution.pdf`

Searching Semi-structured data

- Semi-structured data - is somewhere in between fully structured (tables, excel sheets, databases etc) and unstructured (free text) data.
- Examples: IMDB, Wikipedia entries - although it is user contributed text, there are certain templates, categories etc. There is a relatively uniform formatting. So, it is still possible to uncover some patterns.

Searching Semi-structured data

- Semi-structured data - is somewhere in between fully structured (tables, excel sheets, databases etc) and unstructured (free text) data.
- Examples: IMDB, Wikipedia entries - although it is user contributed text, there are certain templates, categories etc. There is a relatively uniform formatting. So, it is still possible to uncover some patterns.
- Let us say I want to collect the universities where all the presidents of US studied so far from Wikipedia. How should I do that?

Searching Semi-structured data-2

- There are relatively few ways to describe someone's education ("X studied at", "X graduated from", "X has a degree from" etc.)
- If we can come up with a "pattern" that covers all these kinds of sentences, that "pattern" can capture the information we need.

Searching Semi-structured data-2

- There are relatively few ways to describe someone's education ("X studied at", "X graduated from", "X has a degree from" etc.)
- If we can come up with a "pattern" that covers all these kinds of sentences, that "pattern" can capture the information we need.
- Regular expressions are a kind of language to describe such patterns.
- They are used in all programming languages, and even in software such as MS Word (you have to find out how!).
- If you can create a pattern (i.e., I remember my Filename starts with S and has a .pdf extension, but I don't remember its full name) - you can search through lots and lots of text files instantly and get your search results!

Why bother about regular expressions?

- Specifically in the context of search: regular expressions can be used to search through large collections of text corpora (not WWW.. stuff like - parliament proceedings over the years, all writings of Mark Twain etc.)
- Why search through these if there is WWW?

Why bother about regular expressions?

- Specifically in the context of search: regular expressions can be used to search through large collections of text corpora (not WWW.. stuff like - parliament proceedings over the years, all writings of Mark Twain etc.)
- Why search through these if there is WWW?
- We can sometimes have specialized questions (e.g., how many times was "affordable health care" discussed in parliament?) for which web search can be overwhelming and with low what? (precision or recall or both?)

Why bother about regular expressions?

- Specifically in the context of search: regular expressions can be used to search through large collections of text corpora (not WWW.. stuff like - parliament proceedings over the years, all writings of Mark Twain etc.)
- Why search through these if there is WWW?
- We can sometimes have specialized questions (e.g., how many times was "affordable health care" discussed in parliament?) for which web search can be overwhelming and with low what? (precision or recall or both?)

Basic Syntax of Regular Expressions

- searching for ”a” - a
- searching for one or more a’s - a+
- searching for 0 or more a’s - a*
- searching for a or b - a|b
- searching for alphabet, digit, punctuation etc - [: *alpha* :], [: *digit* :], [: *punct* :]
- searching for ”a” at the end of a word. a\b

.... and so on.
See also: `http://www.petefreitag.com/cheatsheets/regex/`

Evaluation of Regular Expressions

- Precision (Correct matches among identified ones)
- Recall (actual number of matches including unidentified ones)

Evaluation of Regular Expressions

- Precision (Correct matches among identified ones)
- Recall (actual number of matches including unidentified ones)
- Example of regular expression usage (in LibreOffice)
- Jargon alert: True positive, True negative, False positive, False negative

Assignment 3 description

- 10 marks, 2 questions (each question has 2 parts, and there is a page 2 for the assignment)
- First question is on Topic 3, Second is on Topic 4
- Due on October 7th
- Description is on Canvas

Next Class

- Lab session with regular expressions exercises

Attendance Exercise

- Consider this passage:
  *This assignment consists of two questions and carries a to- tal of 10 marks. Submit your assignment as a *PDF* file and name it as: your first name–your last name.pdf Late submissions are allowed, but will not be awarded full credit.*
- I want to identify the number of times **s** appeared at the beginning of the word in this. I use the regular expression: \Ws
- What is the precision and what is the recall for this regular expression in terms of achieving its goal?