LING 120, Fall 2017
Language and Computers

Instructor: Sowmya Vajjala

Iowa State University, USA

18 September 2017

Class outline

1. Recap of Week 4
2. Introduction to Search
3. Language and Search

Last class' Exercise

Assume you are given the task of designing a software that can automatically score short answers from students about whatever they read (science, maths, any subject). How will you go about this? What kind of tasks should that system do? What are the features we should look at to evaluate student answers?
a) if we have a target answer
b) if we do not have a target answer

(write your answers on a sheet of paper and return to me. You can also post on Canvas in the discussion forum for today's date)

Your Responses: When we have a target answer

1. Keyword/phrase/n-gram match with target answers (similar to a plagiarism detector!)

Your Responses: When we have a target answer

1. Keyword/phrase/n-gram match with target answers (similar to a plagiarism detector!)
2. There should be large database of semantic relationships between words - so that we can capture words that are not exactly as target answer but similar.

Your Responses: When we have a target answer

1. Keyword/phrase/n-gram match with target answers (similar to a plagiarism detector!)
2. There should be large database of semantic relationships between words - so that we can capture words that are not exactly as target answer but similar.
3. Paraphrasing of target answers should be identified

Your Responses: When we have a target answer

1. Keyword/phrase/n-gram match with target answers (similar to a plagiarism detector!)
2. There should be large database of semantic relationships between words - so that we can capture words that are not exactly as target answer but similar.
3. Paraphrasing of target answers should be identified

Your Responses: When we do not have a target answer

1. Look for words around the words in the question, and compare the words in answers with them.

Your Responses: When we do not have a target answer

1. Look for words around the words in the question, and compare the words in answers with them.
2. Take a probability based approach. If there is no target answer, and there are lot of student answers, we can assume that the frequent answer is the right answer, and go from there.

Your Responses: When we do not have a target answer

1. Look for words around the words in the question, and compare the words in answers with them.
2. Take a probability based approach. If there is no target answer, and there are lot of student answers, we can assume that the frequent answer is the right answer, and go from there.
3. Computer should read the text the student read and figure out answer to the question, and then compare student answer with it.

So how do people solve this?

1. Scoring of maths and science responses, and general content: m-rater, science rater and c-rater by ETS
2. Kaggle-Short Answer Assessment Prize : `https://www.kaggle.com/c/asap-sas`
3. Cognii.com demo `http://cognii.com/demo`

So how do people solve this?

1. Scoring of maths and science responses, and general content: m-rater, science rater and c-rater by ETS
2. Kaggle-Short Answer Assessment Prize : `https://www.kaggle.com/c/asap-sas`
3. Cognii.com demo `http://cognii.com/demo`
4. Going from that question to our next topic:
   - Searching for right answers: `https://www.youtube.com/watch?v=HXIfwL2-4Ek`
   - Asking the right questions: `https://www.youtube.com/watch?v=UIzcIC5RQN8`

Topic 4: Search

1. Introduction to Search (today)
2. Searching through structured vs unstructured data (today)
3. Searching the Web (rest of this week)
4. Searching with regular expressions (next week)
5. Searching through large text corpora (next week)

Warmup questions

1. How many of you use a search engine (google, bing etc)?

Warmup questions

1. How many of you use a search engine (google, bing etc)?
2. How do you use it? (browser/mobile, voice/text etc)

Warmup questions

1. How many of you use a search engine (google, bing etc)?
2. How do you use it? (browser/mobile, voice/text etc)
3. Other than a search engine, where did you have to "search" for information?

Warmup questions

1. How many of you use a search engine (google, bing etc)?
2. How do you use it? (browser/mobile, voice/text etc)
3. Other than a search engine, where did you have to ”search” for information?
4. What kinds of information, in your opinion is easy to search?

Warmup questions

1. How many of you use a search engine (google, bing etc)?
2. How do you use it? (browser/mobile, voice/text etc)
3. Other than a search engine, where did you have to ''search'' for information?
4. What kinds of information, in your opinion is easy to search?
5. How do you search for images or music files?

Warmup questions

1. How many of you use a search engine (google, bing etc)?
2. How do you use it? (browser/mobile, voice/text etc)
3. Other than a search engine, where did you have to ”search” for information?
4. What kinds of information, in your opinion is easy to search?
5. How do you search for images or music files?

Searching is questioning/querying

1. We keep querying, and updating our queries, until we find the results of our search.
2. Note: We also have other possible ways to obtain this information - if you already know where to look for.
3. Search can be related to any form of data (text, speech, image)
4. The data we are searching for can be of different types: structured, unstructured, semi-structured data.

Different types of data

1. Structured - Very organized (e.g., a library database - every book has a title, an author, a publisher, other attributes such as number of pages etc.)
2. Unstructured - free-flowing text from which we should extract what we want (e.g., your typical google search)
3. Semi-structured - where the data is generally unstructured, but there are certain patterns we see, which makes it easy to extract content (e.g., if we want to extract all email addresses from a text).

Searching through Structured Data

1. It is actually quite easy to search through structured data. (Why?)

Searching through Structured Data

1. It is actually quite easy to search through structured data. (Why?)
2. So, what is the problem? Why can't we just work with structured data?

Searching through Structured Data

1. It is actually quite easy to search through structured data. (Why?)
2. So, what is the problem? Why can't we just work with structured data?
3. Have you used lib.iastate.edu to search for books or other stuff before? Is there some structure in the search or is it totally unstructured?

Searching through Structured Data

1. It is actually quite easy to search through structured data. (Why?)
2. So, what is the problem? Why can't we just work with structured data?
3. Have you used lib.iastate.edu to search for books or other stuff before? Is there some structure in the search or is it totally unstructured?
4. On lib.iastate.edu, if I search for "(energy OR effort) AND student success", what does that mean? Is this what we could call structured data? or is it just structured search?

Searching through UnStructured Data

1. What is similar to structured data: Here too, you have to search by some keywords or phrases.
2. Problems: explicit categorization does not exist for text. So, it is not easy to find what you want.

Searching through UnStructured Data

1. What is similar to structured data: Here too, you have to search by some keywords or phrases.
2. Problems: explicit categorization does not exist for text. So, it is not easy to find what you want.
3. It is also very difficult if you have billions of files all over the web, in a WWW search. Should we go through all the files?
4. How do the search engines show search results almost instantly, then?

Evaluating Search Results

1. We know how to evaluate search results based on our information need. How do we decide whether a search system is generally good?

Evaluating Search Results

1. We know how to evaluate search results based on our information need. How do we decide whether a search system is generally good?
2. Relevance: Whether the result a search showed us is actually relevant for the user's need.

Evaluating Search Results

1. We know how to evaluate search results based on our information need. How do we decide whether a search system is generally good?
2. Relevance: Whether the result a search showed us is actually relevant for the user's need.
3. Precision: Of all results returned by the search, how many are actually relevant?
4. Recall: Of all the results that are relevant, how many did the search engine manage to retrieve as relevant?
5. The goal of a good search engine is to provide 100% precision and 100% recall.

Precision and Recall: Competing Priorities

1. How do we achieve 100% recall?

Precision and Recall: Competing Priorities

1. How do we achieve 100% recall?
   ⇒ Just show up everything in the world - that will automatically achieve 100% recall (How??)

Precision and Recall: Competing Priorities

1. How do we achieve 100% recall?
   ⇒ Just show up everything in the world - that will automatically achieve 100% recall (How??)
2. How do we achieve 100% precision?

Precision and Recall: Competing Priorities

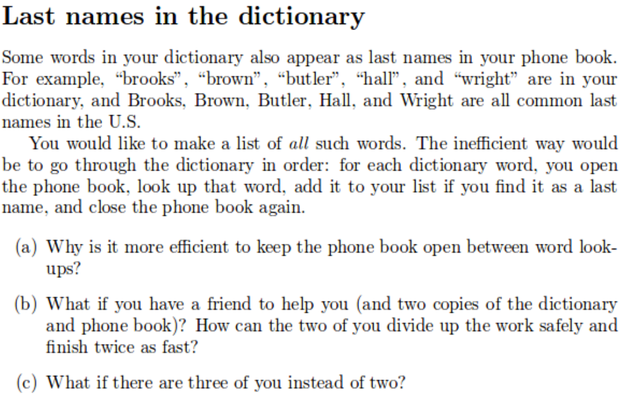
1. How do we achieve 100% recall?
   ⇒ Just show up everything in the world - that will automatically achieve 100% recall (How??)
2. How do we achieve 100% precision?
   ⇒ return that small set of webpages which you are absolutely sure of. (Great, what is the problem then?)

Precision and Recall: Competing Priorities

1. How do we achieve 100% recall?
   ⇒ Just show up everything in the world - that will automatically achieve 100% recall (How??)
2. How do we achieve 100% precision?
   ⇒ return that small set of webpages which you are absolutely sure of. (Great, what is the problem then?)
3. Often, the goal is to reach a balance between precision and recall.

Attendance Exercise: A question about "searching"

Work in groups of 2–3 people, think about a solution for this problem, and return your answers to me giving the names of your team members. You can also submit online on Canvas.



Last names in the dictionary

Some words in your dictionary also appear as last names in your phone book. For example, "brooks", "brown", "butler", "hall", and "wright" are in your dictionary, and Brooks, Brown, Butler, Hall, and Wright are all common last names in the U.S.

You would like to make a list of *all* such words. The inefficient way would be to go through the dictionary in order: for each dictionary word, you open the phone book, look up that word, add it to your list if you find it as a last name, and close the phone book again.

(a) Why is it more efficient to keep the phone book open between word look-ups?

(b) What if you have a friend to help you (and two copies of the dictionary and phone book)? How can the two of you divide up the work safely and finish twice as fast?

(c) What if there are three of you instead of two?

source: `http://nacloweb.org/resources/problems/sample/Phonebook.pdf`