LING 120, Fall 2017
Language and Computers

Instructor: Sowmya Vajjala

Iowa State University, USA

29 September 2017

Class outline

1. Continuing from where we left
2. Review of Topic 4 (by asking more questions!)
3. Time to choose mid-term topics for teams

Grep and Egrep - tools to search large files/folders using regular expressions

- Available by default on Unix based operating systems (Mac, Ubuntu etc)
- Windows 10 perhaps supports it (I don't know - you should tell me!)

Grep and Egrep - tools to search large files/folders using regular expressions

- Available by default on Unix based operating systems (Mac, Ubuntu etc)
- Windows 10 perhaps supports it (I don't know - you should tell me!)
- Brief demo of using egrep:

```
egrep 'Stockmann' pg2446.txt
egrep '[[:digit:]]' pg2446.txt
egrep 'Petra|Horster' pg2446.txt
egrep 'Petra|Horster' pg2446.txt | wc
egrep 'th(e|i)s' pg2446.txt
egrep 'th(e|i)s*' pg2446.txt
```

Regular Expressions: Practice

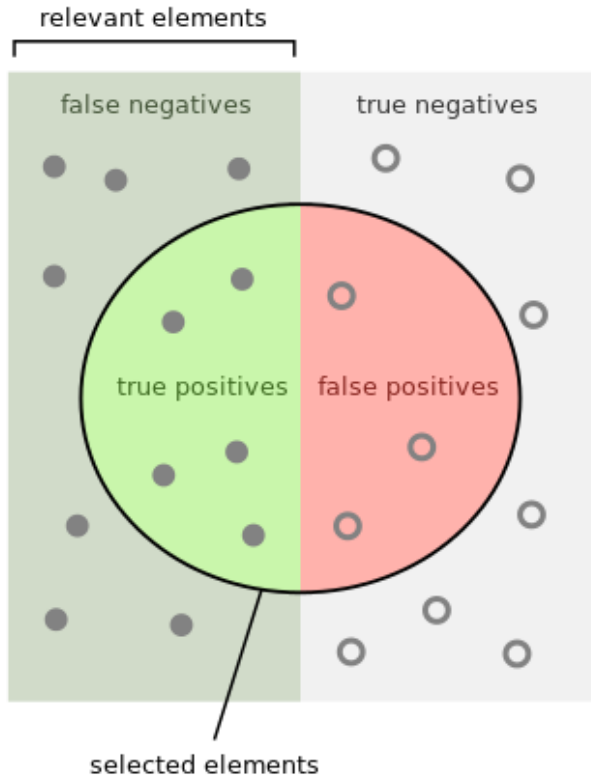
Work in groups of 3 people, download pg2446.txt from `https://goo.gl/gSnhGD`, and answer the following questions using regular expressions.

- Whose name occurs more at the beginning of a line - Dr Stockmann or Mrs Stockmann? How many times?
- What is the difference between *egrep 't.s' pg2446.txt* and *egrep 't.*s' pg2446.txt*?
- In general, what will the pattern *joh?n* identify? How will you verify?
- What will *egrep -o 'Stockmann' pg2446.txt* do? Why? What happens if you remove the -o?
- How many times do Stockmann and Petra appear in the same line in this file?

Note: Regular expressions cheatsheets, egrep help can be found by searching online.

Review: Precision and Recall

source: Wikipedia page on the topic: `https://goo.gl/4gdNRd`



How many selected items are relevant?

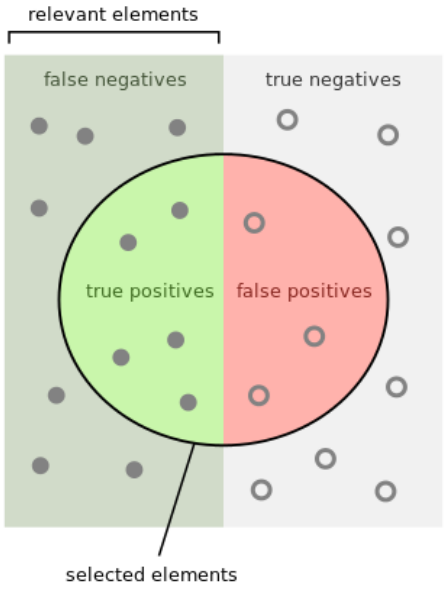
Precision =

How many relevant items are selected?

Recall =

Review: Precision and Recall

source: questions obtained from textbook and quora.com

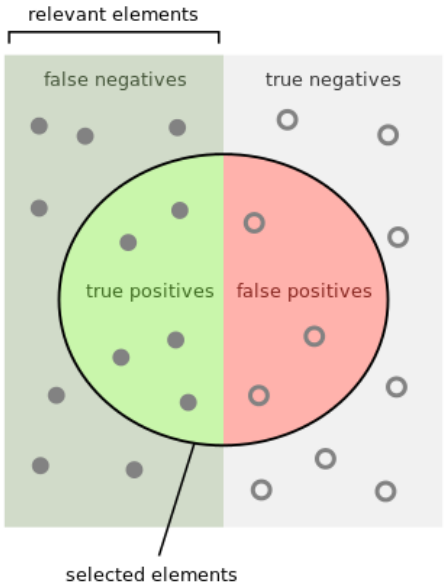


What is important in the following cases - precision or recall?

- Identifying cases where a cancer curing drug has a side effect of nausea

Review: Precision and Recall

source: questions obtained from textbook and quora.com

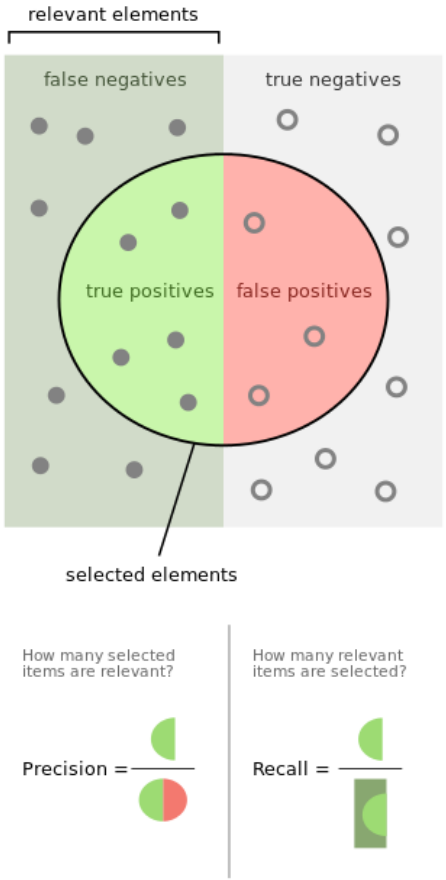


What is important in the following cases - precision or recall?

- Identifying cases where a cancer curing drug has a side effect of nausea
- Identifying cases where a cancer curing drug has a side effect of death

Review: Precision and Recall

source: questions obtained from textbook and quora.com



What is important in the following cases - precision or recall?

- Identifying cases where a cancer curing drug has a side effect of nausea
- Identifying cases where a cancer curing drug has a side effect of death
- Let us say there is a Zombie apocalypse and you are building a safe zone for the non-Zombies. What is a high recall scenario here? What is desirable?

How structured/unstructured is a google search?

- What does: "What does a * do" show me on google?

How structured/unstructured is a google search?

- What does: "What does a * do" show me on google?
- More tips from google on doing such searches:
  - `https://support.google.com/websearch/answer/2466433?hl=en`
  - `https://support.google.com/vault/answer/2474474?hl=en`
- `https://www.google.com/advanced_search?hl=en&fg=1`

How structured/unstructured is a google search?

- What does: "What does a * do" show me on google?
- More tips from google on doing such searches:
  - `https://support.google.com/websearch/answer/2466433?hl=en`
  - `https://support.google.com/vault/answer/2474474?hl=en`
- `https://www.google.com/advanced_search?hl=en&fg=1`
- Why google does not support a full fledged regular expression search

  - `https://www.youtube.com/watch?v=lYiTIDgejas`

How structured/unstructured is a google search?

- What does: "What does a * do" show me on google?
- More tips from google on doing such searches:
  - `https://support.google.com/websearch/answer/2466433?hl=en`
  - `https://support.google.com/vault/answer/2474474?hl=en`
- `https://www.google.com/advanced_search?hl=en&fg=1`
- Why google does not support a full fledged regular expression search
  - `https://www.youtube.com/watch?v=lYiTIDgejas`
- There used to be a Code Search platform by Google which indexed publicly accessible software code - that allowed regular expression search.

Quick recap of topics

- Difference between structured and unstructured search
- How does web-search work?
- How can we use regular expressions to look for patterns?

Quick recap of topics

- Difference between structured and unstructured search
- How does web-search work?
- How can we use regular expressions to look for patterns?
- How is this knowledge useful?: You should now know how the knowledge of patterns and conventions in human language (English) are useful in information extraction.

Quick recap of topics

- Difference between structured and unstructured search
- How does web-search work?
- How can we use regular expressions to look for patterns?
- How is this knowledge useful?: You should now know how the knowledge of patterns and conventions in human language (English) are useful in information extraction.
- If that kind of stuff fascinated you, look for advanced courses on information retrieval, language processing etc (They will ultimately require you to learn to write computer programs!)

Quick recap of topics

- Difference between structured and unstructured search
- How does web-search work?
- How can we use regular expressions to look for patterns?
- How is this knowledge useful?: You should now know how the knowledge of patterns and conventions in human language (English) are useful in information extraction.
- If that kind of stuff fascinated you, look for advanced courses on information retrieval, language processing etc (They will ultimately require you to learn to write computer programs!)
- Professional world: Software engineers, Search Engine Optimizers (yes, that is a title!), Search/Ad relevance evaluators, etc.

Today's attendance exercise

- Group according to your mid-term teams.
- Look at the list of topics and choose one for your team after discussion.
- Start thinking about how to distribute work, what to present etc.
- Return your topic choice along with team member names to me - that is your attendance for today.
- People who are absent today will just not participate in choosing. Don't worry about that.

Next Week

- Introduction to Natural Language Processing - language related issues that trouble a computer
- No assigned readings!
- Start working on Assignment 3
- Start thinking about your mid-term presentation