LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

22 February 2018

Outline

- Review of last class
- Text classification - quick summary of the process (textbook)
- Individual or Group activity - summarizing what we did so far
- Reminder: Tutorial at 5pm, in this room.

Review of last class

Text Classification

- In text classification, we know what the possible categories for our texts can be.
- We also have a collection of texts, already assigned these categories.

Text Classification

- In text classification, we know what the possible categories for our texts can be.
- We also have a collection of texts, already assigned these categories.
- We want to create a "model" of categorization based on this pre-categorized text collection such that the model can "predict" or "assign" categories to new texts.

classification and clustering

- Similarity: we have a collection of texts, we know what are the features we want to look for (words)

classification and clustering

- Similarity: we have a collection of texts, we know what are the features we want to look for (words)
- Difference: In the case of classification, we have a few examples classified into some categories, and we want the machine to learn to classify like this for new texts.

classification and clustering

- Similarity: we have a collection of texts, we know what are the features we want to look for (words)
- Difference: In the case of classification, we have a few examples classified into some categories, and we want the machine to learn to classify like this for new texts.
- In clustering, we do not have such examples, we have no idea how many groupings or possible. We want the machine to figure that out AND do the grouping.

Topic Modeling

- Idea 1: each document is a mixture of topics
- Idea 2: each topic can be represented by groups of words associated with it.

Topic Modeling

- Idea 1: each document is a mixture of topics
- Idea 2: each topic can be represented by groups of words associated with it.
- Idea 3: a word may be very important in one topic, but may not be so important in another topic

Topic Modeling

- Idea 1: each document is a mixture of topics
- Idea 2: each topic can be represented by groups of words associated with it.
- Idea 3: a word may be very important in one topic, but may not be so important in another topic
- So how about looking at a collection of documents, extracting main topics from them and forming clusters of words based on topical similarity?

Assigned Readings from last class

Reading 1

- They discuss the question of automatically identifying the style of writing in newspapers, using text classification
- Data: Four collections (categories) of newspaper articles
- Features: Function words (idea - function word usage is not driven by topic, and hence, truly captures style); POS tag frequencies.
- Learning method: a standard classification algorithm in those days (Ripper)

Reading 1 - What did they find out?

- There are very few features that distinctively identify one category or the other.
- So, what if a given document just does not have those features?
- conclusion: they say they can identify signature features that work for some cases at some times, but not all cases at all time.

Reading 1 - What did they find out?

- There are very few features that distinctively identify one category or the other.
- So, what if a given document just does not have those features?
- conclusion: they say they can identify signature features that work for some cases at some times, but not all cases at all time.
- Note: This is very early work on text classification. It is still a very active area of interest to researchers across disciplines and practitioners in technology industry.

Reading 2

- Machine learning: the method of showing computers a lot of examples of something, so that it can ”learn” from those examples.

Reading 2

- Machine learning: the method of showing computers a lot of examples of something, so that it can "learn" from those examples.
- Classification is a form of machine learning.

Reading 2

- Machine learning: the method of showing computers a lot of examples of something, so that it can "learn" from those examples.
- Classification is a form of machine learning.
- Machine learning is a part of many technologies you use today (email, search, your mobile phone apps, siri/cortana etc)

Reading 2

- Machine learning: the method of showing computers a lot of examples of something, so that it can "learn" from those examples.
- Classification is a form of machine learning.
- Machine learning is a part of many technologies you use today (email, search, your mobile phone apps, siri/cortana etc)
- Applied Linguistics primarily focuses on topics related to teaching, learning languages, instructional settings, assessment by tests etc.
- Machine Learning (specifically text classification) is useful for:

  1. reading/writing/listening/speaking assessment (e.g., in GRE/TOEFL etc)
  2. developing tools that are useful for language learners (e.g., grammarly.com)
  3. doing several other tasks such as educational data mining

Something I came across yesterday

- Title: "Women better represented in Victorian novels than modern, finds study"
- Source: The Guardian (`https://goo.gl/uwkis4`)
- Quick summary - Data: "An analysis of more than 100,000 novels spanning more than 200 years"
- Methods: They used a software called BookNLP, which processes books (.txt files too!), identifies references to characters and groups them, does gender identification (i.e., a form of classification) etc.
- Hypothesis: "expected to see an increase in the prominence of female characters in literature across the two centuries"
- What Data told: "from the 19th century through the early 1960s we see a story of steady decline"

Doing Text Classification in R

(quick summary of chapter 12 in the textbook)

What is the data?

- There is a collection of books written by about 12 authors. 42 books in total.
- There is one anonymous text.
- Task is to "learn" to classify between authors, and use the classification model they learnt to predict the authorship of the anonymous text.

What is the data?

- There is a collection of books written by about 12 authors. 42 books in total.
- There is one anonymous text.
- Task is to "learn" to classify between authors, and use the classification model they learnt to predict the authorship of the anonymous text.
- But there are only 42 texts for all authors together, which is not really suitable for teaching a machine to learn to classify.
- Text classification typically needs 100s of examples per category.
- How did he handle this issue?: he decided to split each file into 10 equal parts, and treat each part as if it is a text in itself. (i.e., 420 texts now!)

... continued

- So, each of these 420 texts is represented as word-frequency tables before proceeeding to next step.

... continued

- So, each of these 420 texts is represented as word-frequency tables before proceeeding to next step.
- Once we have this, next step is to build a term document matrix

... continued

- So, each of these 420 texts is represented as word-frequency tables before proceeeding to next step.
- Once we have this, next step is to build a term document matrix
- He does "bag of words" classification i.e., each word is a potential "signature feature".

… continued

- So, each of these 420 texts is represented as word-frequency tables before proceeeding to next step.
- Once we have this, next step is to build a term document matrix
- He does "bag of words" classification i.e., each word is a potential "signature feature".
- If we take all words, it will be too many and the "learning" becomes slow. So, he talks about removing words that don't appear frequently enough to be useful.
- Once all this is done, the the actual "learning" part is only one line of R code!
- Once done, that "learned model" can be used for predicting the authorship of the unknown texts.

Rest of Today's class

- I created small functions to do all the corpus analyses we learnt so far, and stored them in a R file.
- There is a small tutorial associated. It has exercises at the end.
- Go through the tutorial - line by line - try to understand what is happening. Perhaps add comments to yourself.
- Once you understand whats going on, start doing the exercises.
- We will primarily continue on this in the evening tutorial (if anyone comes)
- Ask questions, and I suggest working in groups of 2 and discussing with your teammate.

Next Week

- movie reviews sentiment analysis using tm library.
- no readings. Just spend some time revising what we learnt so far - it will be useful.