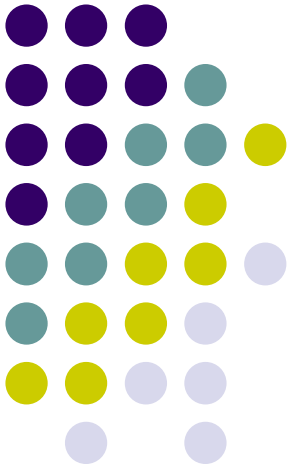
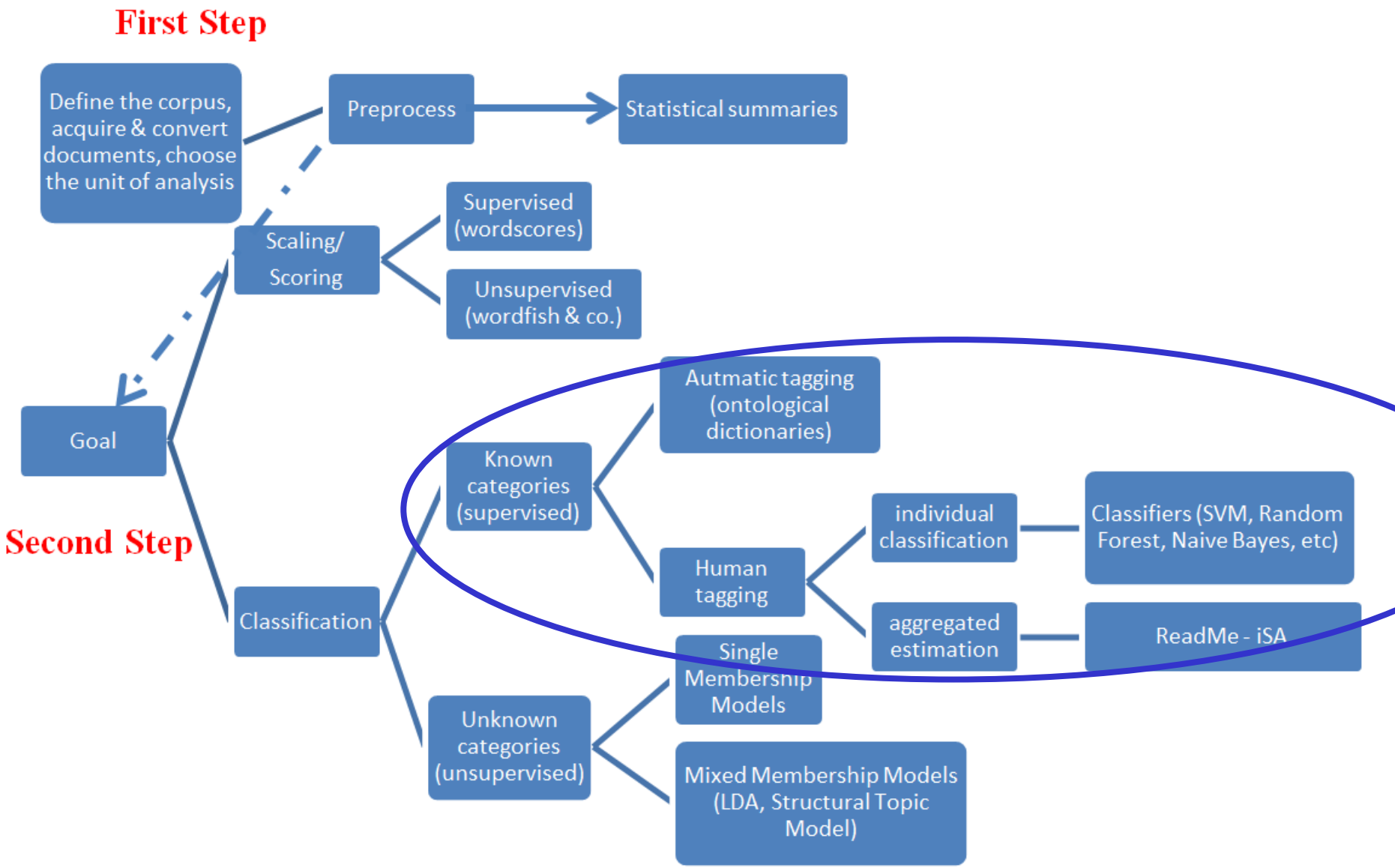
Applied Scaling & Classification Techniques in Political Science

Lecture 5 – Part 2

Supervised classification methods: automatic tagging



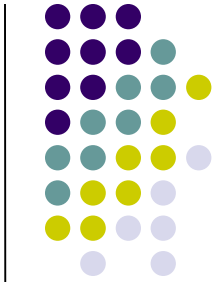
Our Course Map



Reference

- ✓ Grimmer, Justin, and Stewart, Brandon M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3): 267-297

Classification methods



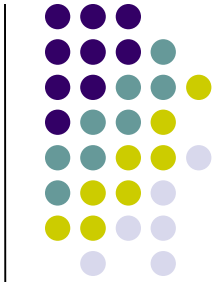
Classifying Documents into Known Categories

Assigning texts to **some known categories** is the most common use of content analysis methods in political science

For example, researchers may ask if local news coverage is positive or negative, if legislation is about the environment or some other issue area, if international statements are belligerent or peaceful, etc.

In each instance, the goal is to infer **either the category of each document**, the **overall distribution of documents across categories**, or **both**

Classification methods



Human-based methods for making these inferences are both time and resource intensive

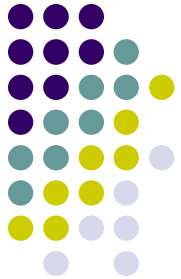
Automated methods can mitigate the cost of assigning documents to categories

There are **two broad groups of supervised classification methods** available according to the type of **tagging** (i.e., the assignation of a document to a given pre-defined category) **employed**:

We can have either

1) automatic tagging - *dictionaries*
2) human tagging **-** *supervised learning methods*

Human tagging

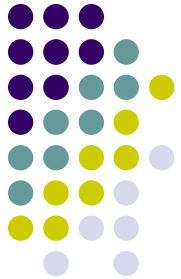


Supervised learning methods replicate the familiar manual coding task, but with a machine

First, human coders are used to classify a subset of documents into a predetermined categorization scheme

Then, this training set is used to train an automated method, which then classifies the remaining documents

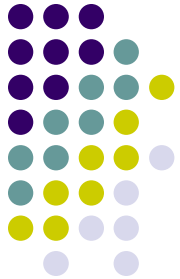
Automatic tagging



Dictionaries use the **relative rate** at which key words appear in a text to **classify documents into categories** or **to measure the extent to which** documents belong to particular categories

Let’s start with automatic tagging…

Dictionary methods



Suppose the goal is to measure the **tone** (also called the "**sentiment"**) in newspaper articles: whether articles convey information positively or negatively about a given topic

A **dictionary to measure sentiment** is a list of words that are either dichotomously classified as positive («good», «fantastic», etc.) or negative («bad», «horrible», etc.) or contain more continuous measures of their content

You can then use that dictionary to identify **the tone of a document**: either positive or negative according to the relative number of words in that document identified by the dictionary as positive or negative ones

Dictionary methods

Formally, within a given dictionary Z each word *m* (m=1,….M) will have an associated score s_m

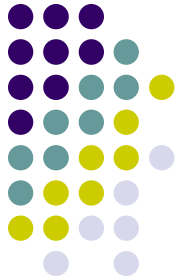
For the simplest measures, s_m =-1 if the word is associated with a negative sentiment and s_m =+1 if associated with a positive sentiment

The analyst then applies some decision rule, such as summing over all the weighted feature values, to create a score for the document

For example, if $N_i = \sum_{m=1}^{M} W_{im}$ words included in dictionary Z are used in document *i*, then dictionary methods can use such list of words to measure the sentiment for any document t_i as:

$$t_i = \sum_{m=1}^{M} \frac{s_m W_{im}}{N_i}$$

Dictionary methods



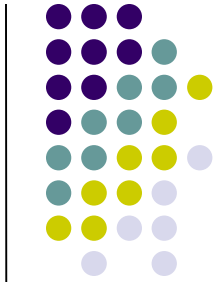
Scholars often use t_i as an approximately continuous measure of document sentiment, that is, it allows us to sort documents as to which are more or less positive or negative relative to one other

t_i can also be used to classify documents into **sentiment categories** if a decision rule that identifies a cut point is assumed along with the dictionary method

Perhaps the simplest coding rule would assign all documents with $t_i > 0$ to a positive sentiment category and $t_i < 0$ to a negative sentiment

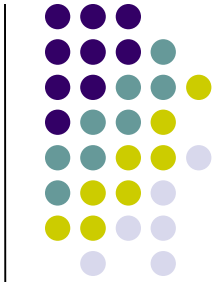
And if $t_i = 0$? Either neutral category or NC

Dictionary methods



Of course, the words included in the texts you are analyzing that are **not also included** in the dictionary, will not provide any additional information for your classification aim (we will discuss more about this point later on)

Dictionary methods

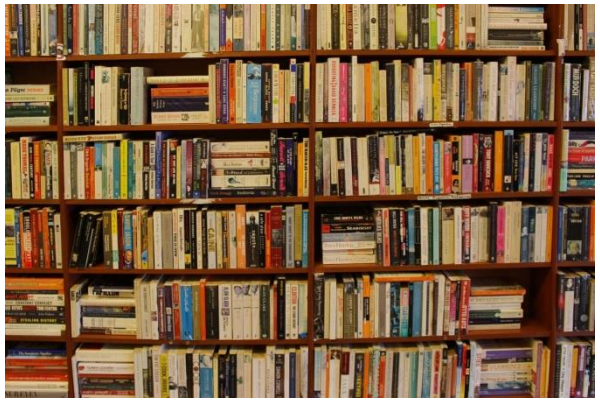


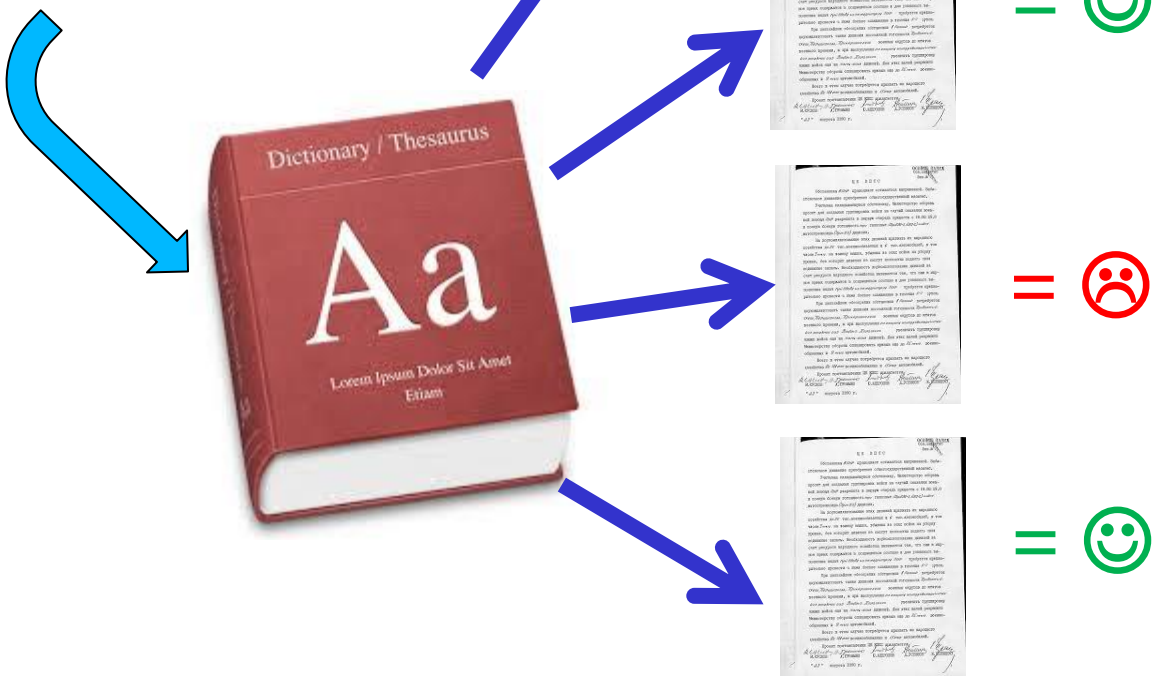
Sentiment analysis is just one type of analysis a dictionary method can perform

The general idea of dictionaries make them relatively easy and cheap to apply across a variety of problems: identify **words that separate categories** (for example *policy categories*) and measure **how often those words occur in texts**

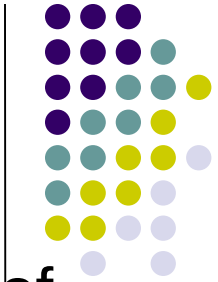
For example, the Lexicoder Topic Dictionary (Albugh et al., 2013) contains 1,387 keywords under 28 topics (e.g., macroeconomics, civil rights, health care, agriculture) based on the Comparative Agenda Project's coding scheme. If you are interested about it, just let me know!

Dictionary methods





Dictionary methods

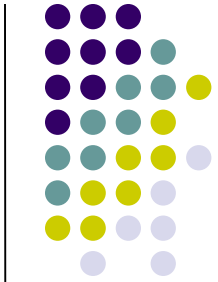


Using a dictionary can therefore **minimize** the amount of labor needed to classify documents (and this is attractive!) BUT…

…the difficulty lies in constructing a dictionary so that all relevant terms are included (**no false negatives**, i.e., terms we should have included in the dictionary cause they are relevant given our research topic, but failed to do so), but no irrelevant or wrong terms are (**no false positives**, i.e., terms we have included in the dictionary but should not have, being them irrelevant given our research topic)

This is not that easy!

Dictionary methods



…the challenges of using a dictionary:

For dictionary methods to work well, the scores attached to words must closely align with **how** the words are used in a particular context

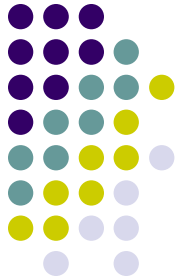
If a dictionary is developed for a **specific application**, then this assumption should be easy to justify

But when dictionaries are **created in one substantive area and then applied to another**, serious errors can occur

Language do **change across topics**!

For example, a word like `cancer` may have a positive connotation in a health-care company, but negative in many other contexts

Dictionary methods



Moreover, dictionary methods work pretty well when you study texts that use a **standardized language** (i.e., legal text!).

In other contexts, things become more complex…

Dictionary methods

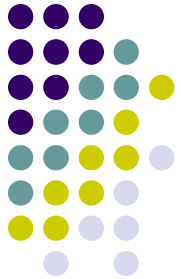
…language in fact evolves continuously: one cannot code all possible semantic rules (double meaning sentences, specific jargons, neologisms, irony) unless reading the posts!!!

Dictionary methods

On the other side, counting the number of positive and negative terms in a sentence may lead to **paradoxical effects**

Dictionary methods

Dictionary methods



Dictionaries, therefore, should be used with **substantial caution**, or at least coupled with **explicit validation**

The problem is that quite often measures from dictionaries are **rarely validated**

Rather, standard practice in using dictionaries is to assume the measures created from a dictionary are correct and then apply them to the problem

The consequence of **domain specificity and lack of validation** is that most analyses based on dictionaries are built on shaky foundations

Dictionary methods

If using dictionaries, choose therefore a dictionary **appropriate** to the task at hand, and **validate** the utility of the dictionary, for example by confirming that a sample of dictionary-generated scores of text in the corpus conform to human coding of the text for the measure of interest

R pakcages to install

install.packages("gridExtra", repos='http://cran.us.r-project.org')
install.packages("syuzhet", repos='http://cran.us.r-project.org')
install.packages("plotly", repos='http://cran.us.r-project.org')
install.packages ("reshape2", ='http://cran.us.r-project.org')
install.packages ("wordcloud", ='http://cran.us.r-project.org')
install.packages ("tm", ='http://cran.us.r-project.org')
install.packages ("plyr", ='http://cran.us.r-project.org')