Text Mining for the Social Sciences
Lecture 1: Introduction

Stephen Hansen

Text as Data

Text data is a sequence of characters.

The typical dataset is made up of a collection (corpus) of documents.

We can think of each document as an observation.

Text data is *unstructured*: the information we typically want is not immediately accessible.

Text mining studies how to provide an informative, quantitative representation of documents.

This representation could either be an end in itself, or else the first step in a broader statistical study.

Example

Doc1 = [‘text’, ‘mining’, ‘is’, ‘more’, ‘fun’, ‘than’, ‘coal’, ‘mining’]

Examples of representations:

1. Does the document contain 'coal'? In this case, Doc1 = True (or = 1).
2. Bag of words:

text	mining	is	more	fun	than	coal
1	2	1	1	1	1	1

Example

Doc1 = [‘text’, ‘mining’, ‘is’, ‘more’, ‘fun’, ‘than’, ‘coal’, ‘mining’]

Examples of representations:

1. Does the document contain ’coal’? In this case, Doc1 = True (or = 1).
2. Bag of words:

text	mining	is	more	fun	than	coal
1	2	1	1	1	1	1

Note that the bag of words representation is the same as for the document

Doc2 = [‘coal’, ‘mining’, ‘is’, ‘more’, ‘fun’, ‘than’, ‘text’, ‘mining’]

Does this matter?

Relevant Information

Suppose we want to identify documents about text mining. The bag of words representation is useful.

Now suppose we want to predict the missing word. The bag of words representation is less useful.

$$\text{Doc1} = [\text{`text'}, \text{`mining'}, \text{`is'}, \text{`more'}, \text{`fun'}, \text{`than'}, \text{`coal'}, ?]$$

We could instead use bigram counts

(text, mining)	(mining, is)	(is, more)	(more, fun)	...
1	1	1	1	...

Note that we still lose information.

Main Message

Whether or not a particular representation is useful cannot be separated from the context for which it's being used.

Any meaningful representation of text will throw away some information; that's the whole point of text mining in the first place.

Knowing which information to keep and which to throw away is the art of text mining (and of data science more generally).

Why "for the Social Sciences"?

I chose the title of the course intentionally.

Much of text mining and information retrieval has been developed by computer scientists.

Important questions include: what are efficient data structures for holding bags of words?, how to process massive corpora like Wikipedia?, what aspects of text mining can be parallelized?, etc.

These are very important questions in some domains, like search engine optimization.

They are less crucial (but certainly not irrelevant) for social science research:

1. Many interesting datasets are not particularly "Big".
2. Much more emphasis on the question being asked with the data.
3. Can seek out specialized help when necessary.

Unsupervised versus Supervised Learning

This course will mostly be concerned with finding interesting structure in unlabeled documents (unsupervised learning).

Another branch of text mining seeks to predict document labels using text features (supervised learning). Examples: sentiment analysis, Gentzkow and Shapiro (2010).

A section of Topics in Big Data Analytics, which follows this course, will cover supervised learning applications.

Structure of Course

Broadly speaking, the course will be divided as follows:

1. Preliminaries: acquiring and processing text data.
2. Count-based analysis: Boolean methods, dictionary methods, tf-idf weighting, vector space model
3. Latent variable analysis: Latent Semantic Analysis, multinomial mixture modeling, Latent Dirichlet Allocation
4. Variational inference

Throughout, there will be a mix of statistical ideas, programming examples, and real-world applications.

Evaluation

Evaluation will come from two sources:

1. Four homeworks, mix of programming and theory, 10% each.
2. Project due on last day of term, Friday 12 June, 60%.

Details of both to follow. No final exam for course.

Textbooks

The following two books will cover the statistical ideas for the course:

1. Manning, Raghavan, and Schütze (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

2. Murphy (2012). *Machine Learning: a Probabilistic Perspective*. MIT Press.

A free copy is available of the first, and other machine learning textbooks' content overlaps considerably with the second.

Programming Language

The programming language we will use for the course is Python, whereas up until now you have used R.

The statement "language X is better than language Y" is meaningless. Better question: "I tend to do tasks like Z. Does X or Y let me do task Z more efficiently?"

So why Python?

Advantages of Python

Python is a general-purpose scripting language:

1. Elegant, readable syntax.
2. Very popular, large user base.
3. Nice scientific computing stack (numpy, scipy, ipython, matplotlib).
4. Used in many applications besides scientific computing (Youtube, Dropbox, etc.).

Python provides a unified tool for acquiring text data, handling unstructured strings, and statistical analysis.

Main disadvantage with respect to R: less developed statistical libraries.

Also has reputation as slow with respect to compiled languages. More on this if we have time.

Getting Started with Python

I taught myself Python from `http://learnpythonthehardway.org/book/`.

Rishabh gave an introductory session, materials available.

For scientific computing in particular, I recommend McKinney (2012). *Python for Data Analysis*. O'Reilly.

Getting Started with Python

I taught myself Python from `http://learnpythonthehardway.org/book/`.

Rishabh gave an introductory session, materials available.

For scientific computing in particular, I recommend McKinney (2012). *Python for Data Analysis*. O'Reilly.

Note that a default Python installation does not include the scientific computing packages.

If you are brave, you can try hand-building these. I would recommend installing the Anaconda distribution instead.

Note distinction between Python 2 and 3.

My Way into Text Data

I have long-standing interest in behavior of monetary policymakers.

Initial papers on voting behavior in committees. Votes are easy to quantify.

Most time in committees spent deliberating, but almost no research on this in economics.

Joined with Michael McMahon and Andrea Prat to start research in this area.

First output: "Transparency and Communication on the FOMC: A Computational Linguistics Approach"

The Data

A large amount of text data available from http://www.federalreserve.gov/monetarypolicy/fomc_historical.htm.

We first focused on FOMC transcripts from the era of Alan Greenspan. 149 meetings from August 1987 through January 2006.

A document is a single statement by a speaker in a meeting.

There are 46,502 such statements.

Associated metadata: speaker biographical information, macroeconomic conditions, etc.

No memory problems: data is around 40 MB file.

The Challenge

There are 5,594,280 total alphanumeric tokens in the data; 25,725 unique tokens.

Consider the bag of words representation in terms of a *document-term matrix*, whose (d, v)th element is the count of token v in document d.

What are the characteristic features of this matrix?

What is the "Big Data" problem relevant for this context?

Spirit of Course

I view my primary value as putting together various ideas in such a way that will allow you to productively apply text mining to applications in economics and finance (academic, regulatory, commercial).

Please feel free to share your expertise during class, everyone will benefit!