Coursera Capstone: Task 0

Leigh Matthews

December 11, 2017

Task 0: Natural Language Processing.

NLP tools and techniques help businesses process, analyze, and understand all of this data in order to operate effectively and proactively.

The first step in analyzing any new data set is figuring out: (a) what data you have and (b) what are the standard tools and models used for that type of data. Make sure you have downloaded the data from Coursera before heading for the exercises. This exercise uses the files named LOCALE.blogs.txt where LOCALE is the each of the four locales *en_US, de_DE, ru_RU and fi_FI*. The data is from a corpus called HC Corpora (*www.corpora.heliohost.org*). See the readme file at http://www.corpora.heliohost.org/aboutcorpus.html for details on the corpora available. The files have been language filtered but may still contain some foreign text.

In this capstone we will be applying data science in the area of natural language processing. As a first step toward working on this project, you should familiarize yourself with Natural Language Processing, Text Mining, and the associated tools in R. Here are some resources that may be helpful to you. - Natural language processing Wikipedia page - Text mining infrastucture in R - CRAN Task View: Natural Language Processing - Coursera course on NLP (not in R)

Dataset

This is the training data to start that will be the basis for most of the capstone. You must download the data from the Coursera site and not from external websites to start (**https://www.coursera.org/learn/data-science-project/supplement/Iimbd/task-0-understanding-the-problem**). Capstone Dataset

Your original exploration of the data and modeling steps will be performed on this data set. Later, if you find additional data sets that may be useful for building your model you may use them.

Tasks to accomplish 1.Obtaining the data - Can you download the data and load/manipulate it in R? 2.Familiarizing yourself with NLP and text mining - Learn about the basics of natural language processing and how it relates to the data science process you have learned in the Data Science Specialization.

Getting the Data

CAUTION: The data files are very large, so be ware of possible issues with working memory.

Set up the R Studio environment, set the working directory, and load required packages.

```
options(warn = FALSE)  # Remove warnings
options(digits=2)      # Set to 2 decimal places
#setwd("H:/Capstone")
library(knitr); library(tm); library(pander); library(ascii); library(rmarkdown)
```

Check to see if the downloaded file already exists; if not, download and unzip the raw data.

Load the data files into R.

```
suppressWarnings(news.raw <- readLines("H:/Capstone/Data/en_US.news.txt"))
suppressWarnings(twitter.raw <- readLines("H:/Capstone/Data/en_US.twitter.txt"))
suppressWarnings(blogs.raw   <- readLines("H:/Capstone/Data/en_US.blogs.txt"))
```

Basic Text Data Information:

Note that the file sizes are very large and thus computing memory needs to be considered. Below is a summary table for the data files, showing the siZe of the file, the length (number of lines), and the maximum number of characters per line.

http://stackoverflow.com/questions/15488350/programmatically-creating-markdown-tables-in-r-with-knitr

Source	File Size	Length (# of Lines)	Max Char per Line
"Blogs"	248.5 Mb	899288	40835
"News"	19.2 Mb	77259	5760
"Twitter"	301.4 Mb	2360148	213

Questions to consider

1.What do the data look like?

The corpora data is formatted as text files, which vary in size from 19.2 Mb to 301.4 Mb. Each file contains thousands of lines of text.

2.Where do the data come from?

The corpora are collected from publicly available sources by a web crawlerthat checks for language. Each entry is tagged with it's date of publication. Where user comments are included they will be tagged with the date of the main entry.

Each entry is tagged with the type of entry, based on the type of website it is collected from (e.g. newspaper or personal blog). If possible, each entry is tagged with one or more subjects based on the title or keywords of the entry, and if a tag is not feasible, the entry is tagged with a '0'.

To save space, the subject and type is given as a numerical code. Once the raw corpus has been collected, it is parsed to remove duplicate entries and split into individual lines. (Source: Coursera About the Corpora)

3.Can you think of any other data sources that might help you in this project?

No data sets in particular stand out to to help with this project.

4.What are the common steps in natural language processing?

According to Paul Nelson (https://insidebigdata.com/2017/07/10/five-steps-tackling-big-data-natural-language-processing/), the five major steps of NLP are below.

STEP 1: BASIC PROCESSING Content with the most important information is written down in a natural language such as English, Spanish, etc., and it is not conveniently tagged. Therefore, to extract information from this content you need to do text mining, text extraction, or full-up natural language processing. The input to natural language processing will be a simple stream of Unicode characters (typically UTF 8), and basic processing is required to convert this character stream into words, phrases, and syntactic markers which can then be used to better understand the content. Basic processing includes language identification, sentence detection, lemmatization, decompounding, structure extraction, tokenization, entity and phrase extraction.

STEP 2: IDENTIGY LEVEL OF UNDERSTANDING AND EVALUATE FEASIBILITY Next, decide what level of content understanding is required - macro vs. micro. While micro understanding (extracts understanding from individual phrases or sentences) generally contributes to macro understanding (provides a general understanding of the document as a whole), the two can be entirely different.
You should also evaluate the project feasibility as not all NLP understanding projects are possible within a reasonable cost and time. Ask questions like: What are the accuracy requirements? Can you afford the time and effort? Is the text short or long? And, is a human involved? If you decide it's feasible to move forward then it's time to extract the content.

STEP 3: EXTRACT CONTENT FOR MACRO/MICRO UNDERSTANDING Begin with macro understanding to perform functions such as: . Classifying / categorizing / organizing records . Clustering records . Extracting topics . Keyword / key phrase extraction . Duplicate and near-duplicate detection . Semantic search If you need to understand individual words and phrases, use micro understanding for the extracting of individual entities, facts or relationships from the text. This is useful for: . Extracting acronyms and their definitions . Extracting key entities like people, company, product, location, dates, etc. Remember that micro understanding must be done with syntactic analysis of the text - this means that order and word usage are important.

STEP 4: MAINTAIN TRACEABILITY Acquiring content from multiple sources and extracting information from that content will likely involve many steps and a large number of computational stages. Thus, it is vital to provide traceability for all outputs. You can then trace back through the system to identify exactly how that information came to be, supporting quality analysis and validation purposes. Note the following: . The original web pages which provided the content . The start and end character positions of all blocks of extracted text . The start and end character positions for all entities, plus the entity IDs . Cleansing or normalization functions applied / used by all content

STEP 5: INCORPORATE HUMAN FEEDBACK Content understanding can never be complete without some human intervention. You need a human to discover new patterns and for creating, cleansing or choosing lists of known entities, to name a few. Many of these processes can be mind-numbingly repetitive. In a large-scale system, you will need to consider the human element and build that into your NLP system architecture. Be aware that continuously doing quality analysis of the data during each step of the process is key to getting the best understanding of natural language content. The whole process may seem daunting, but using these steps and techniques as a guide can help you create a working, robust system for acquiring, harvesting, and turning unstructured big data into practical, insightful knowledge that advances your use case.

Text analysis processes

Subtasks-components of a larger text-analytics effort-typically include: . Information retrieval or identification of a corpus is a preparatory step: collecting or identifying a set of textual materials, on the Web or held in a file system, database, or content corpus manager, for analysis. . Although some text analytics systems apply exclusively advanced statistical methods, many others apply more extensive natural language processing, such as part of speech tagging, syntactic parsing, and other types of linguistic analysis. . Named entity recognition is the use of gazetteers or statistical techniques to identify named text features: people, organizations, place names, stock ticker symbols, certain abbreviations, etc . Disambiguation-the use of contextual clues-may be required to decide where, for instance, "Ford" can refer to a former U.S. president, a vehicle manufacturer, a movie star, a river crossing, or some other entity. . Recognition of Pattern Identified Entities: Features such as telephone numbers, e-mail addresses, quantities (with units) can be discerned via regular expression or other pattern matches. . Coreference: identification of noun phrases and other terms that refer to the same object. . Relationship, fact, and event Extraction: identification of associations among entities and other information in text . Sentiment analysis involves discerning subjective (as opposed to factual) material and extracting various forms of attitudinal information: sentiment, opinion, mood, and emotion. Text analytics techniques are helpful in analyzing, sentiment at the entity, concept, or topic level and in distinguishing opinion holder and opinion object. . Quantitative text analysis is a set of techniques stemming from the social sciences where either a human judge or a computer extracts semantic or grammatical relationships

between words in order to find out the meaning or stylistic patterns of, usually, a casual personal text for the purpose of psychological profiling etc.

5.What are some common issues in the analysis of text data?

Common issues in NLP inlcude the ambiguity problem, language variability, specific nouns (i.e. Fruit fly) being interpreted as phrases, spam detection, Part of Speech tagging, named entity recognition, sentiment analysis, coreference resolution, word sense disambiguation, parsing, machine translation, information translation, and text summarization.

6.What is the relationship between NLP and the concepts you have learned in the Specialization?

Natural Language Processing is an extension of Data Mining.