Assignment 1: Unix

L555

Due Tuesday, September 6

1. **N-grams**: Change the commands to create a list of bigrams (from the in-class exercises) so that it creates trigrams instead of bigrams from the file *dates_in_may.txt*. Send me the commands in one file and the output in another.

2. **Rhyme sorting**: Create a wordlist out of *dates_in_may.txt* that is sorted alphabetically, but right-to-left, i.e. `tina` should appear before `angst`. Hint: the command `rev` takes a file as argument and reverses the characters per line. Send me the commands in one file and the output in another.

3. **Debugging**: The following commands are supposed to take the file *vm.pos* and make one file with all the words that are common nouns (NN), one file for proper nouns (NP), and one file with all the 3rd person singular verbs (VBZ). Then these files are to be packed and zipped. Unfortunately, the student who wrote this made 3 mistakes. Find them and correct them.

```
grep NN vm.pos > vm.nouns
grep NP vm.pos > vm.nouns
grep VBZ vm.pos > vm.verbs
tar xvzf vm.nouns vm.names vm.verbs > vm.tgz
```

4. **Mystery**: What does the following command line do?

```
cat data/dates_in_may.txt | tr "aeiou" "X" | tr -sc "X" "\n" | grep XX | wc
```

5. **Frequency gathering**: Automatically create a "dictionary" of POS tags from *vm.pos*, i.e. a list that has one POS tag per line plus their frequency. Assume that all upper case characters belong to POS tags, and lower case characters to words (and can thus be deleted).

6. L555 only: **Data cleaning**: Go to Project Gutenberg (`http://www.gutenberg.org`) and download a text file of a book.

   (a) Using various Unix commands, obtain the plain text without header or footer information and without meta-information contained in the text itself (e.g., `[Sidenote: A RUINED HEDGE]`), so that you have only the text itself.

   (b) Using Unix commands, obtain a frequency dictionary of words in their lower-case forms.

7. Nothing to turn in: **Install software**: Installing software is something you'll want to become more comfortable with. Rejoice if everything goes right; don't panic if it doesn't; search the internet if you get unfamiliar error messages.

   Task: make sure that you have the following installed on your machine and/or have access to them:

   (a) Python 3.2 or higher: `https://www.python.org/download/`

   (b) NLTK: `http://www.nltk.org/install.html`

   - If you're feeling groovy, you could consider installing Anaconda Python, which also installs a number of Python packages: `https://store.continuum.io/cshop/anaconda/`
   - You should also at this time explore various tools for writing & running programs in Python (e.g., IDEs: `https://wiki.python.org/moin/IntegratedDevelopmentEnvironments`)

   Once you've done this, test your installation: run `python3` in a terminal or command prompt and then, **within python**, type `import nltk`

   Once here, install the NLTK data (`http://www.nltk.org/data.html`). Namely, type:

```
import nltk
nltk.download()
```

When the downloader comes up, select *All packages* and make sure that the Download Directory is specified to be for "central installation" (e.g., `/usr/share/nltk_data` for mac & unix users). If you see a permission error, exit python and then type (for mac/unix users): `sudo python3`. This will give you admin rights to install the data in a shared folder. (See also suggestions on the website.) ... This step will take some time.

- You may also have to set the environment variable `NLTK_DATA`, to make sure NLTK knows where its data is. To do this involves creating or modifying something like a `.profile` or `.bash_profile` file in your home directory.
- Type: `ls -al ~` to see if you have such a file already.
- Either modify or create the file, such that you have a line which says something like: `export NLTK_DATA=/usr/share/nltk_data` (verifying that this is indeed where the NLTK data is stored)
  - You can check where it's checking by typing within Python: `import nltk`, followed by: `nltk.data.path`