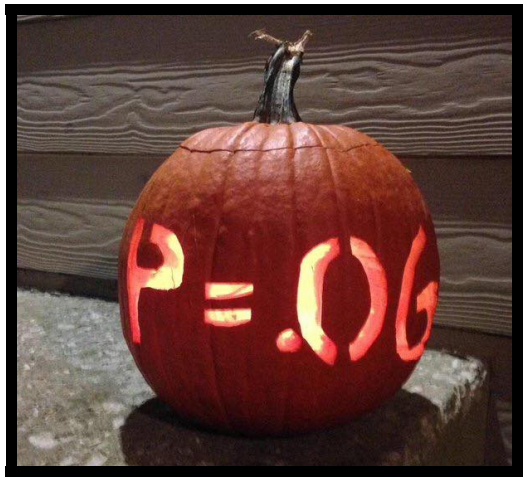
An Overview of Text Analysis Tools

Rachael K. Hinkle

University at Buffalo, SUNY

October 26, 2017



Some Prelimaries

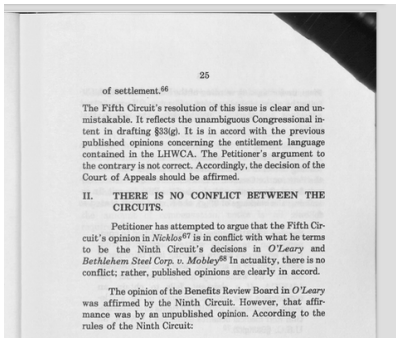
- You can download these slides and all relevant files at:
  - `http://rachaelkhinkle.com/TextStuff.zip`
  - Files won't be online for long
- If you would like to follow along you will need:
  - R
  - iMacros (browser plug-in) for Firefox
  - Python, NLTK, Jupyter (all included in Anaconda)

Outline of Session

- Overview of Tools for:
  - Gathering Data
  - Processing Textual Data
  - Creating Text-based Variables for Analysis
- Tips and Tricks
- Discussion
- Optional Demos

Tools for Gathering Textual Data

- API
- RSS feeds
- Web Scraping
  - **iMacros**
  - Python
  - R
- Old-fashioned Manual Labor
  - Download by hand
  - Scan/photograph hard copies



25

of settlement.[66]

The Fifth Circuit's resolution of this issue is clear and unmistakable. It reflects the unambiguous Congressional intent in drafting §33(g). It is in accord with the previous published opinions concerning the entitlement language contained in the LHWCA. The Petitioner's argument to the contrary is not correct. Accordingly, the decision of the Court of Appeals should be affirmed.

II. THERE IS NO CONFLICT BETWEEN THE CIRCUITS.

Petitioner has attempted to argue that the Fifth Circuit's opinion in *Nicklos*[67] is in conflict with what he terms to be the Ninth Circuit's decisions in *O'Leary* and *Bethlehem Steel Corp. v. Mobley*[68] In actuality, there is no conflict; rather, published opinions are clearly in accord.

The opinion of the Benefits Review Board in *O'Leary* was affirmed by the Ninth Circuit. However, that affirmance was by an unpublished opinion. According to the rules of the Ninth Circuit:

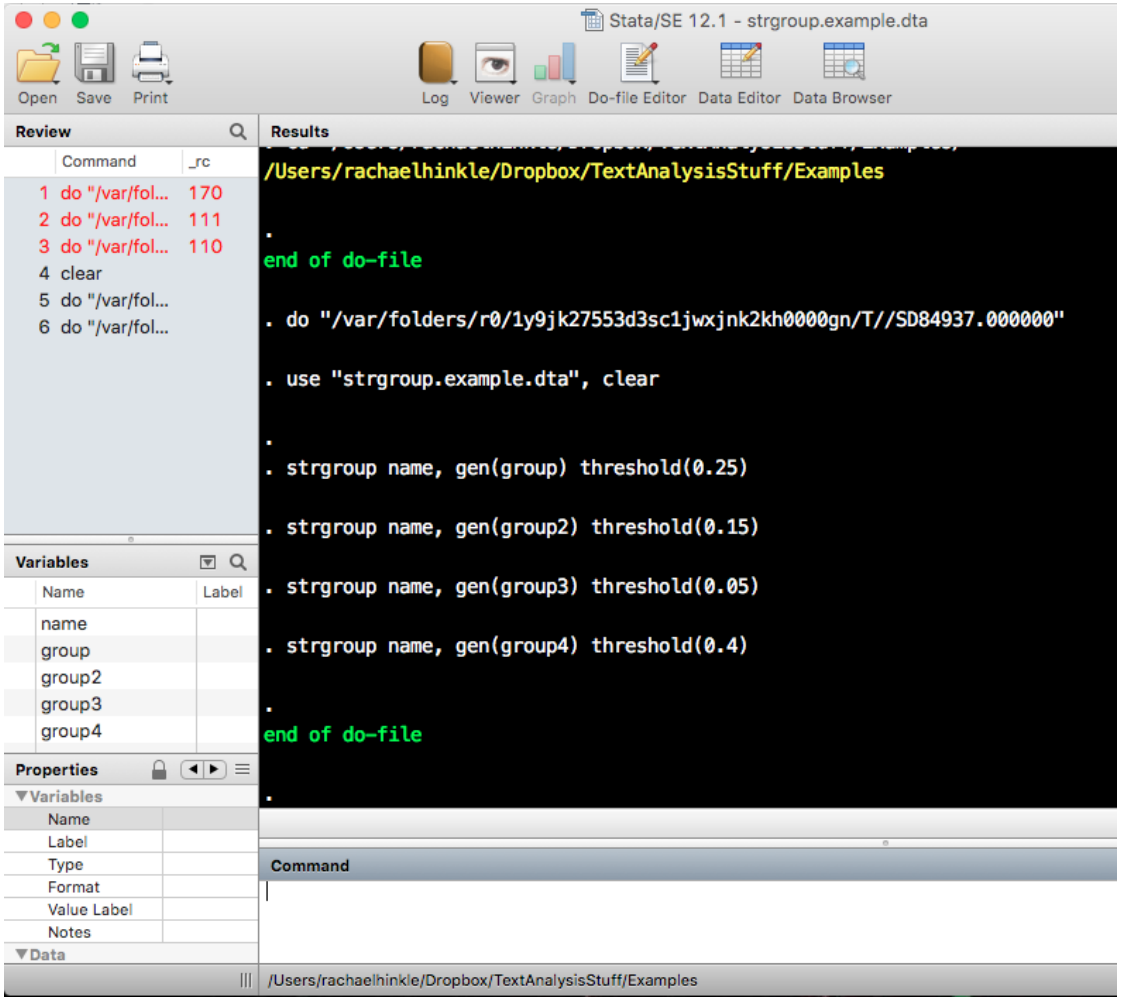
HOW TO: Using iMacros

- Install plug-in for your browser.
- Complete one iteration of a task by hand.
- Write code to loop through list of tasks.
- Figure out why the code is not working correctly.
- Watch your computer collect information automatically.

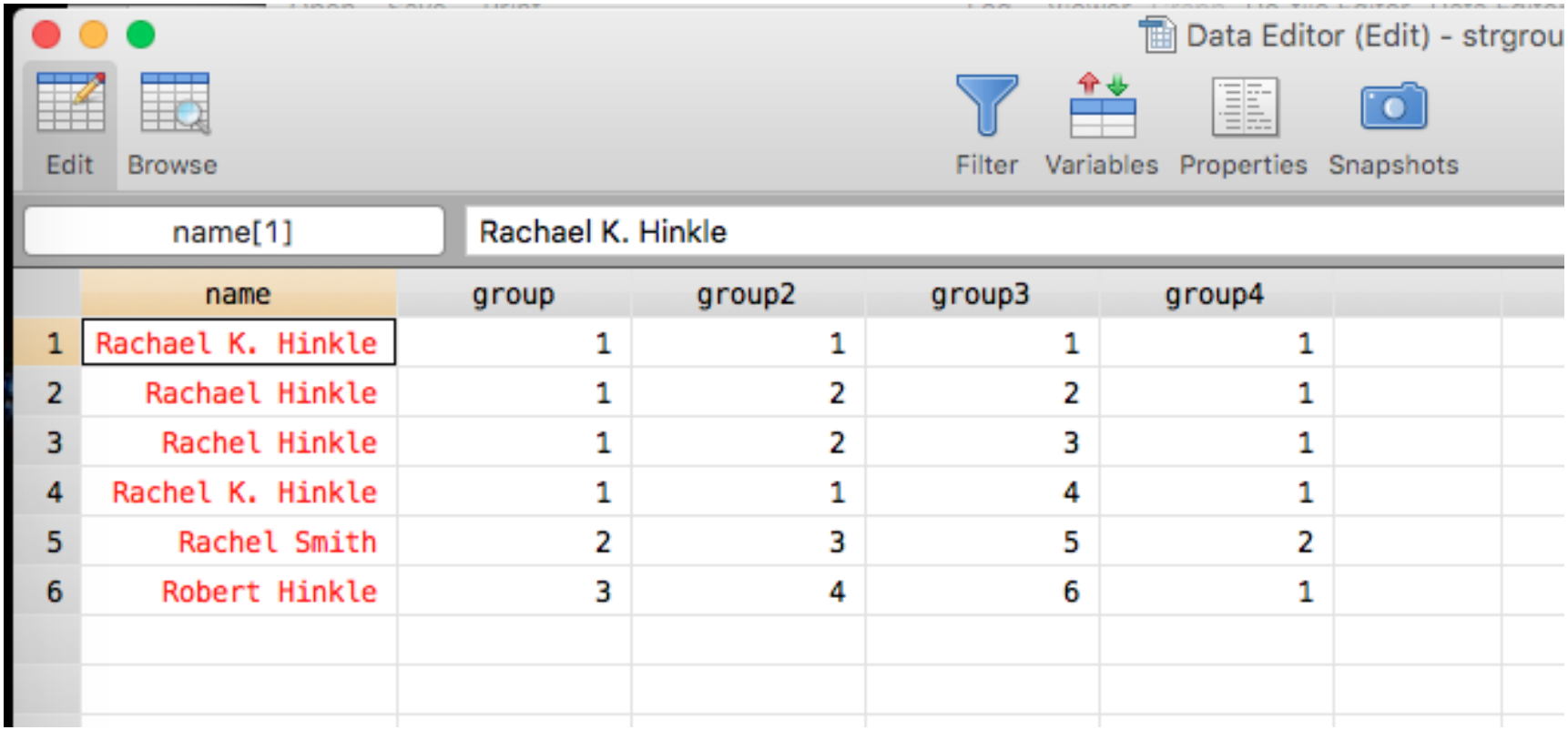
Tools for Processing Textual Data

- OCR
  - This works surprisingly well, although not free
  - ABBYY FineReader ($65-$85)
  - Adobe Acrobat Pro ($89, free(?))
  - PDFelement 6 Pro ($99)
- Python, NLTK
  - Input each file in a folder; output each file in a new folder
  - Input each file in a folder; output .csv file
  - Use regular expressions to extract chunks of text
- Stata
  - Use regular expressions: regexr, regexm
  - Identify similar strings: **strgroup**
- Excel
  - Filtering
  - Functions: Upper, Trim, Concatenate, Hyperlink

Example: strgroup



Example: strgroup



	name	group	group2	group3	group4
1	Rachael K. Hinkle	1	1	1	1
2	Rachael Hinkle	1	2	2	1
3	Rachel Hinkle	1	2	3	1
4	Rachel K. Hinkle	1	1	4	1
5	Rachel Smith	2	3	5	2
6	Robert Hinkle	3	4	6	1

HOW TO: Integrating Human and Computer Resouces

First Step: Identify documents to be excluded

1. **Python**: Extract unwieldy chunks of text, export to .csv
2. **Excel**: Apply filter to text chunk to code
3. **Human**: Read text chunk to code variable by hand
4. **Excel**: Create hyperlink to files for uncoded rows
5. **Human**: Click link and read file to code variable by hand

Second Step: Delete Irrelevant Documents

6. **Excel**: Use CONCATENATE function to write code
7. **Terminal/Command**: Change directory, "cd"
8. **Terminal/Command**: Run code from Excel, "rm"/"del"

Creating Text-based Variables for Analysis

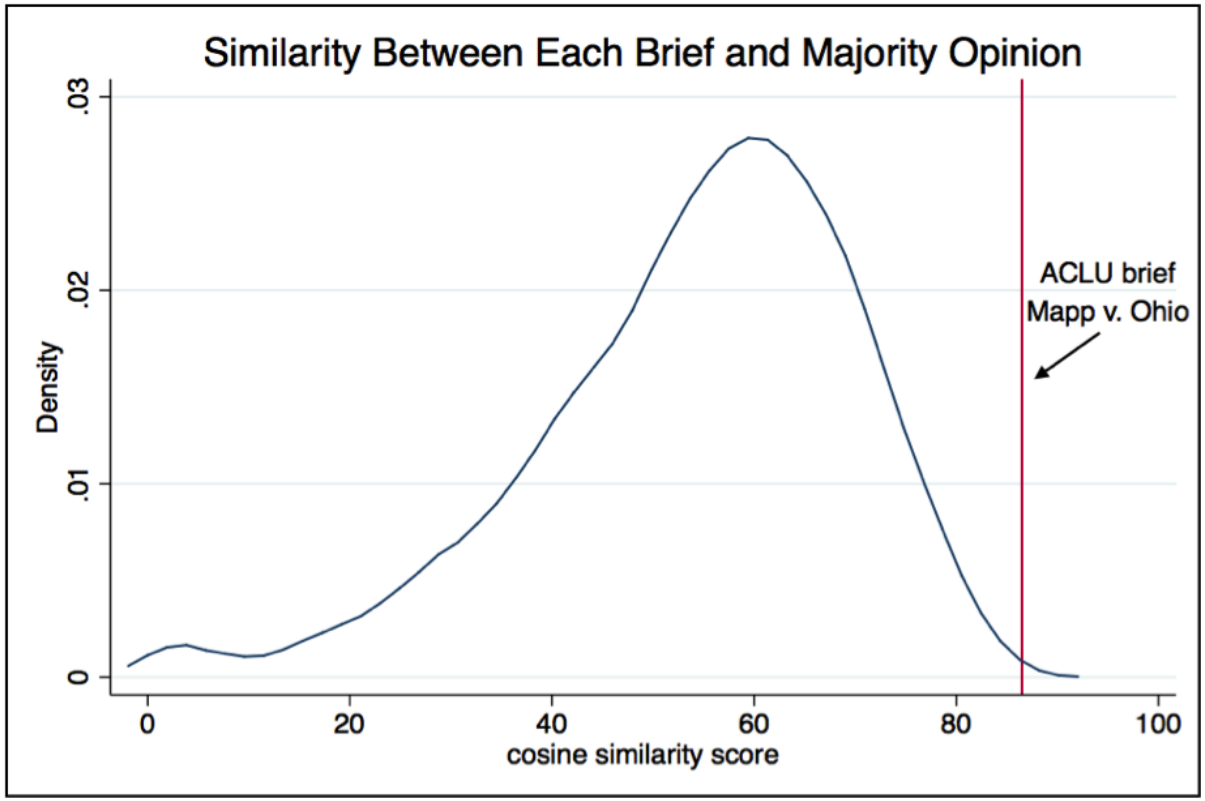
Characteristics of a Document

- Number of words
- Size of vocabulary
- Frequency of some domain-relevant token
- Summary of the sentiment of text
  - LIWC ($90)
  - Vader
  - ProfilerPlus
- Readability
- Political position: Wordfish (R)
- Number of issues or topics: TextTiling (NLTK)
- Any hand-coded classification: Supervised topic models

Dyadic Measures for Pairs of Documents

- **WCopyFind**: Common phrasing
  - Percentage of Doc A that appears in Doc B.
  - Percentage of Doc B that appears in Doc A.
- Cosine Similarity: Similarity between words used in two documents
  - Can be implemented in Python.
  - Word importance is weighted using entire corpus.
- Python: Anything you can think of
  - % of precedents cited in Doc A that appear in Doc B.
  - # of words or phrases in Doc A that appear in Doc B.
  - # of proper nouns common to two documents.

Cosine Similarity



HOW TO: Using WCopyfind

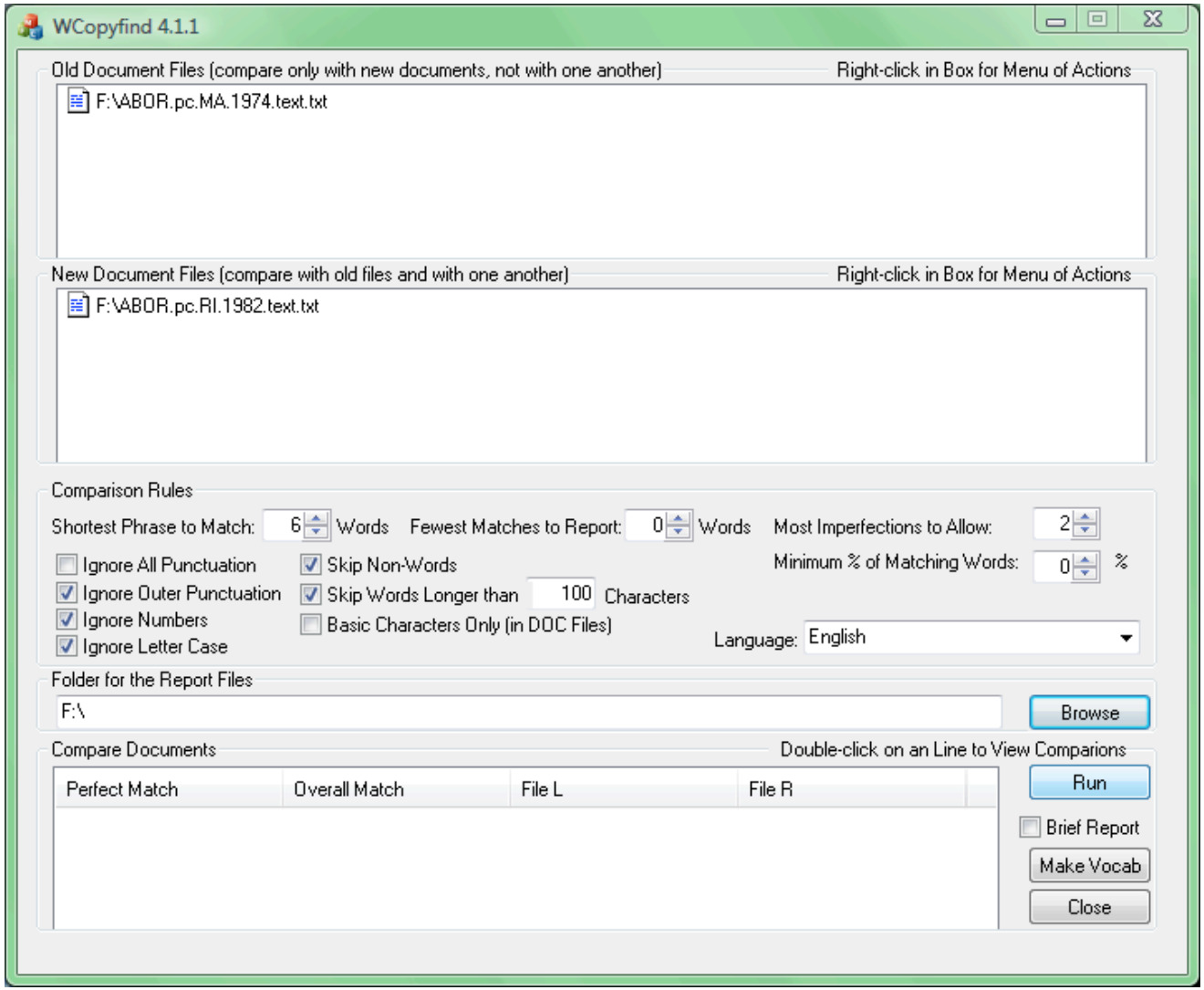
- Download from The Plagiarism Resource Site
- Can run the .exe file on a Mac:


Wine

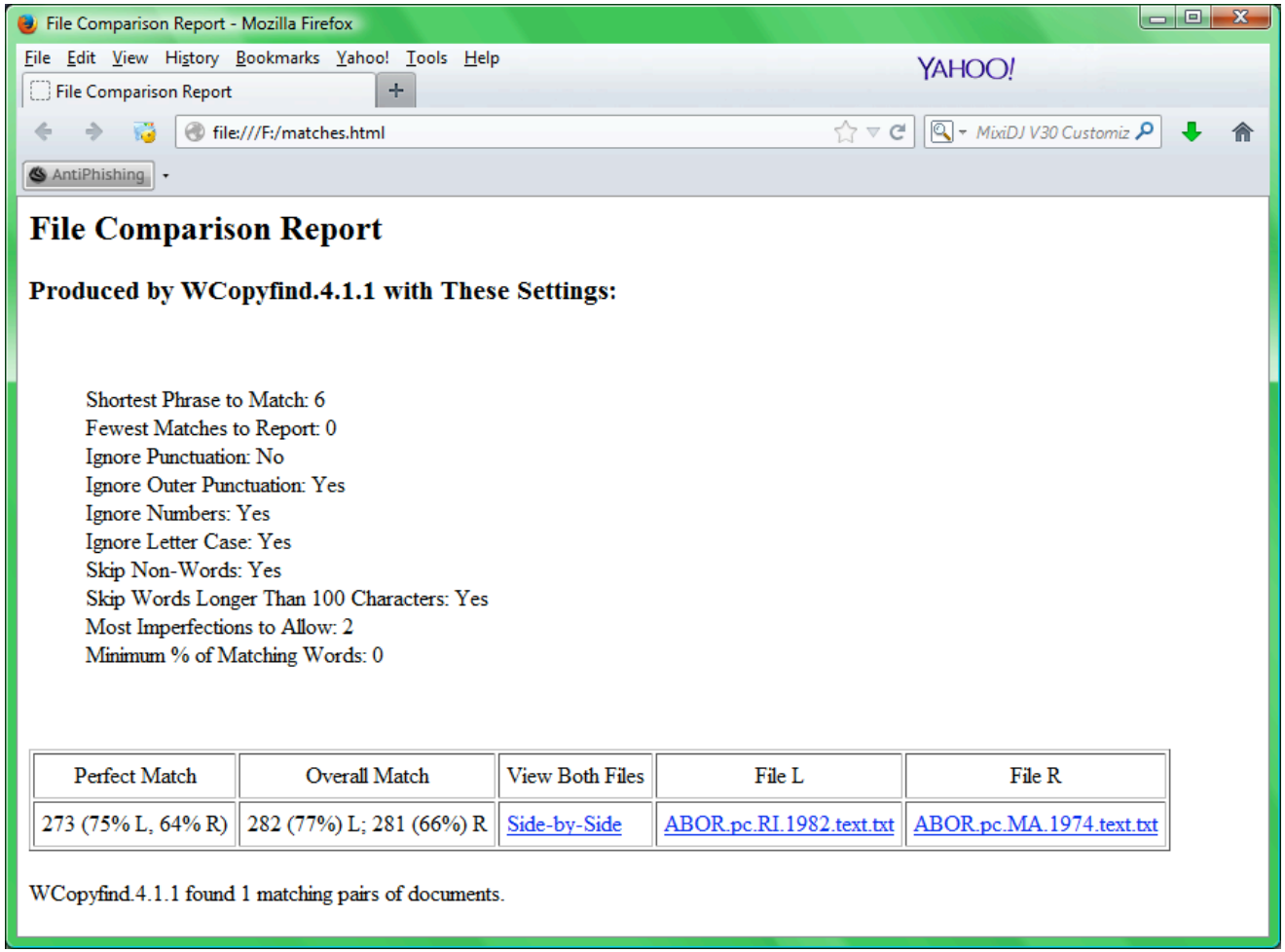

WineBottler

- Input: Two batches of documents
- Select options: Literature has examples to follow
- Output: Detailed and Overall Summary
- Processing Output: Can copy to Excel to process

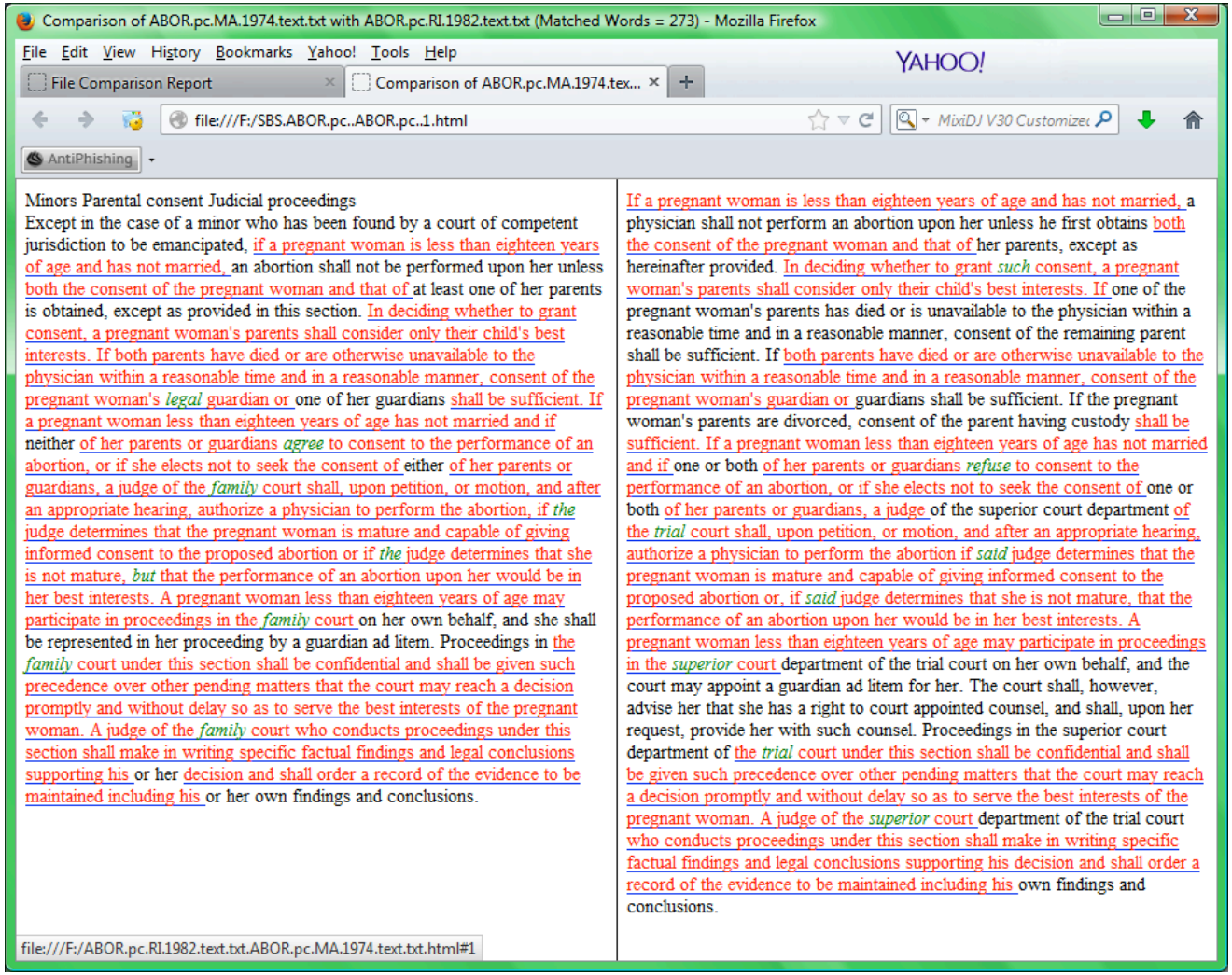
WCopyfind: Input and Options



WCopyfind: General Output

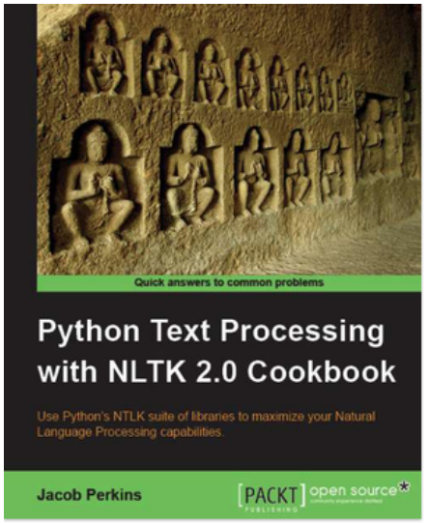


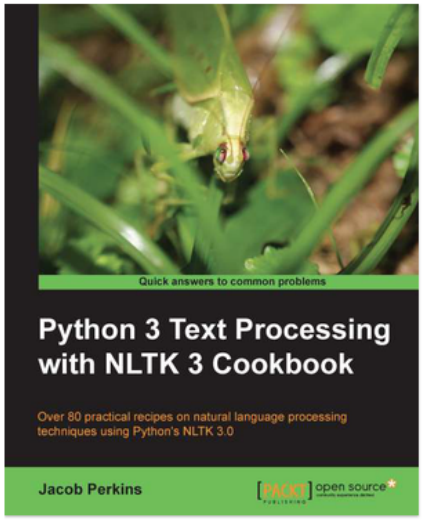
WCopyfind: Detailed Output: Side-by-Side View



Tips and Tricks - Specific

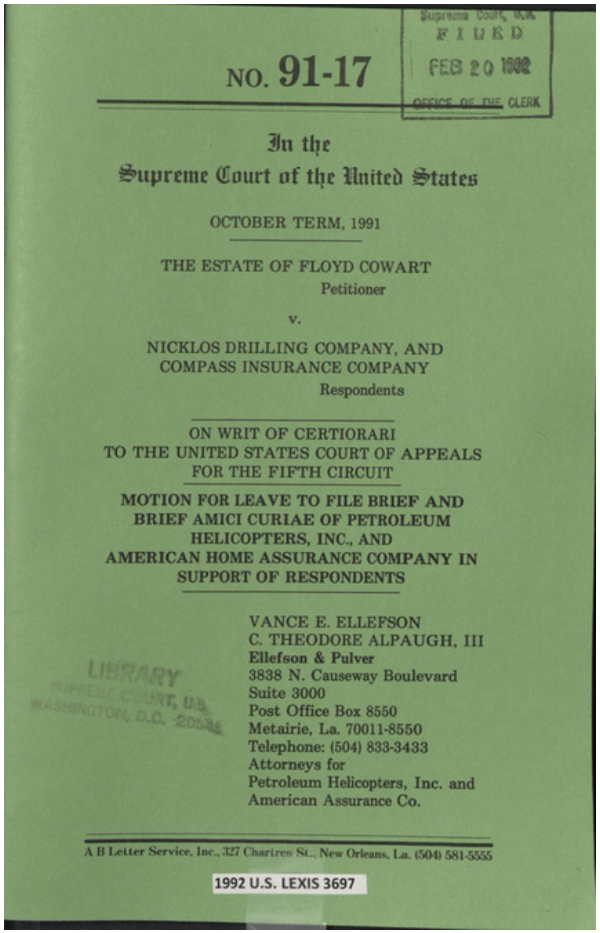
- Spaces in filenames and filepaths make life complicated
  - snake_case_filenames_work
  - iPreferCamelCase
- You can run scripts without knowing how to write scripts
  - Mac: Automater
  - Windows: Action(s)
- Jacob Perkins's books from Packt Publishing are great





Tips and Tricks - General

- Never underestimate the power of a low-tech hack.
- Plan in advance when you will move on.
- Don't lose sight of the theory.
- Don't overcomplicate things.



NO. 91-17

FILED
FEB 20 1992
OFFICE OF THE CLERK

In the
Supreme Court of the United States

OCTOBER TERM, 1991

THE ESTATE OF FLOYD COWART
Petitioner

v.

NICKLOS DRILLING COMPANY, AND
COMPASS INSURANCE COMPANY
Respondents

ON WRIT OF CERTIORARI
TO THE UNITED STATES COURT OF APPEALS
FOR THE FIFTH CIRCUIT

MOTION FOR LEAVE TO FILE BRIEF AND
BRIEF AMICI CURIAE OF PETROLEUM
HELICOPTERS, INC., AND
AMERICAN HOME ASSURANCE COMPANY IN
SUPPORT OF RESPONDENTS

VANCE E. ELLEFSON
C. THEODORE ALPAUGH, III
Ellefson & Pulver
3838 N. Causeway Boulevard
Suite 3000
Post Office Box 8550
Metairie, La. 70011-8550
Telephone: (504) 833-3433
Attorneys for
Petroleum Helicopters, Inc. and
American Assurance Co.

A B Letter Service, Inc., 327 Chartres St., New Orleans, La. (504) 581-5555

1992 U.S. LEXIS 3697

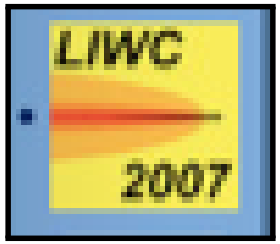
Discussion

- How can these tools be used in your research?
- What questions do you have about specific data?

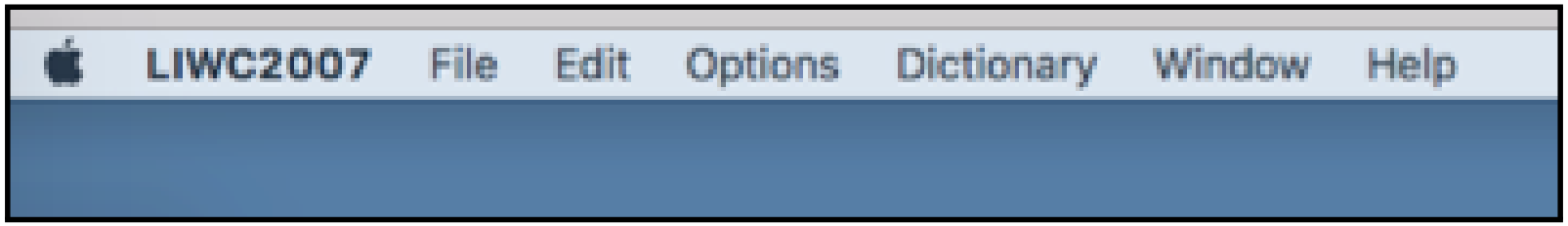
Linguistic Inquiry Word Count

- The $90 is worth it if you do any work with text.
- Make sure the dictionaries make sense in your domain.
- Making your own dictionary is quite simple.
- Options for what to include in report.
- Easy to run on a folder full of documents.
- Output is % of words in document from that category.
- Categories can overlap.

LIWC: Simple Interface

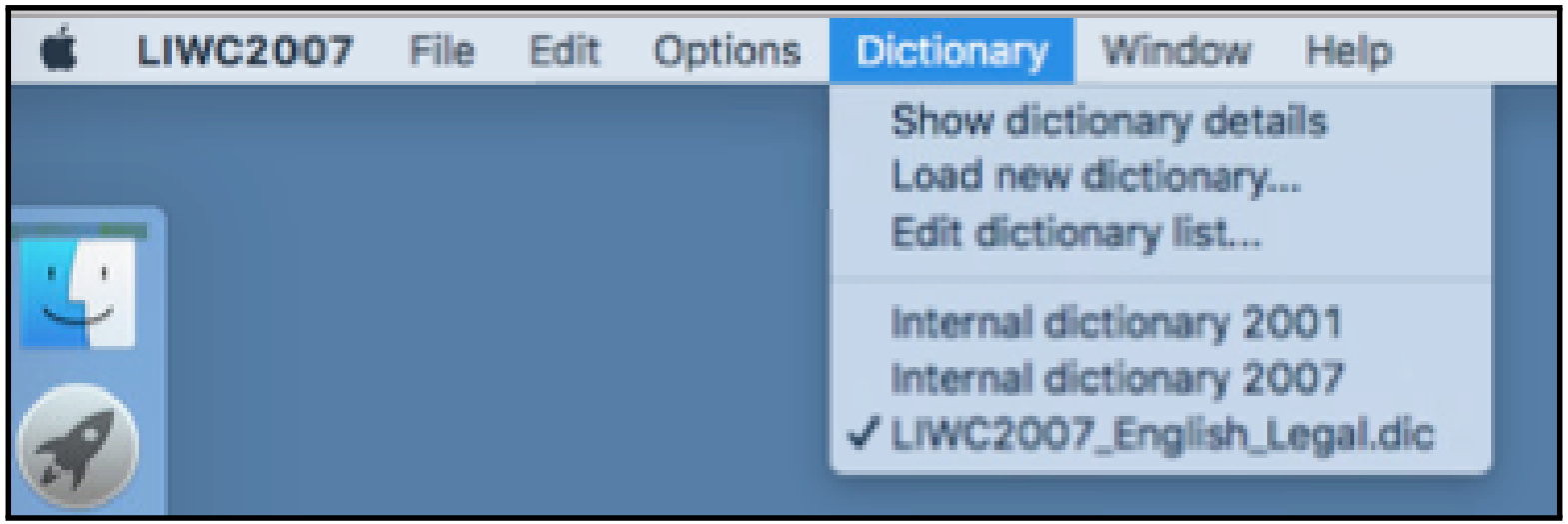


Desktop Icon



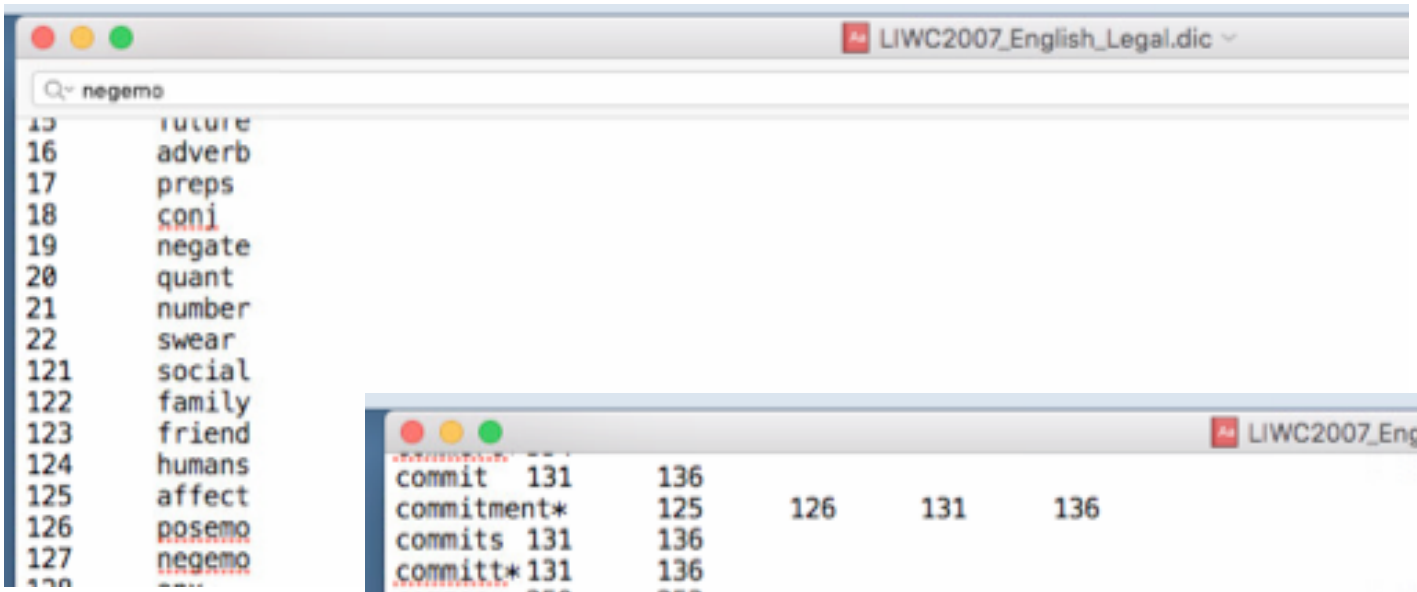
Menu

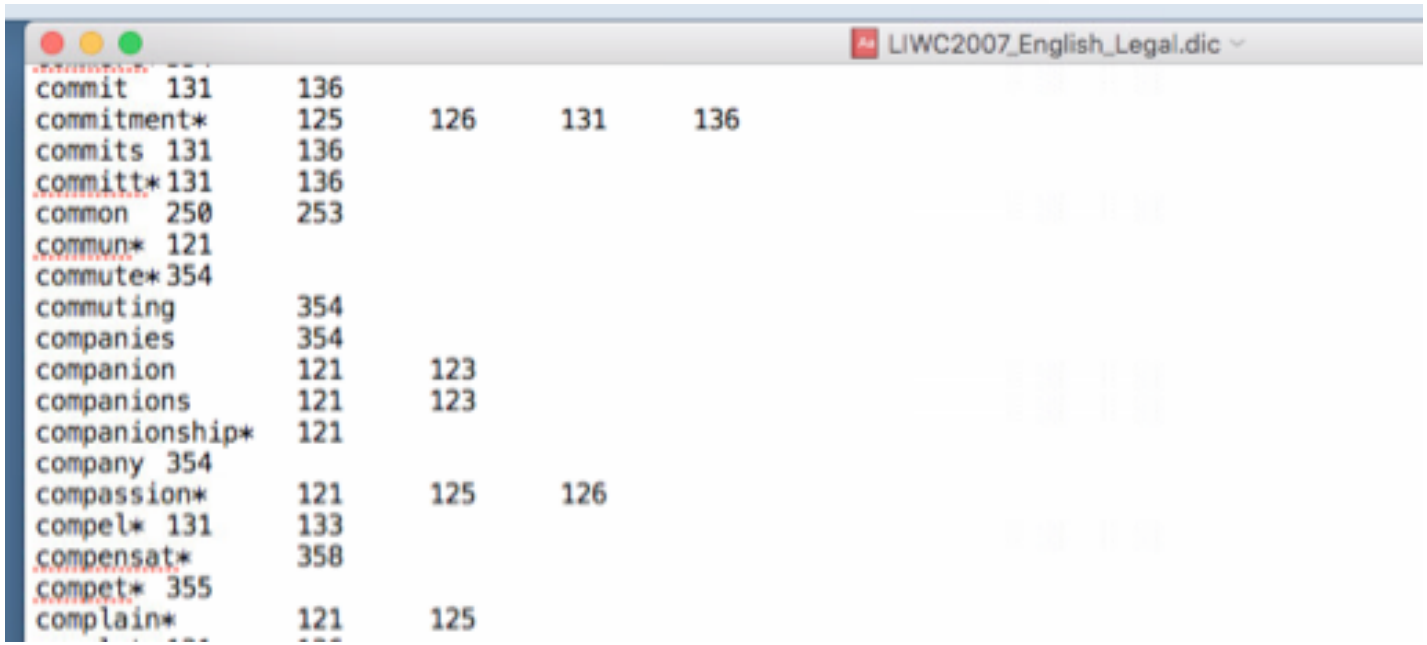
LIWC: Using the Right Dictionary



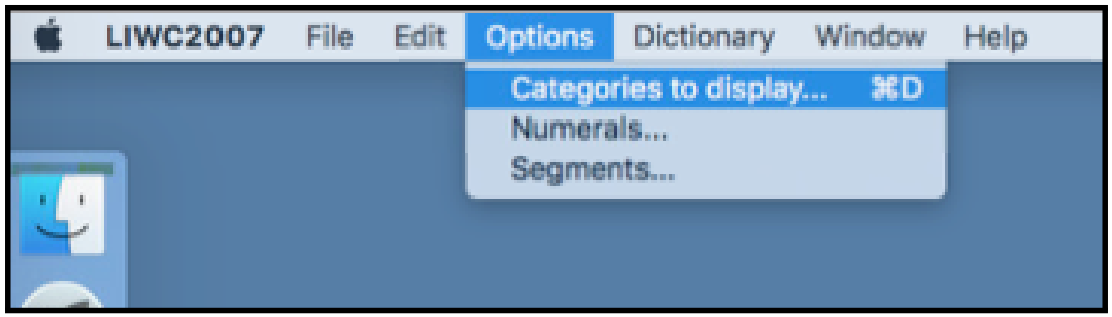
Select Dictionary File

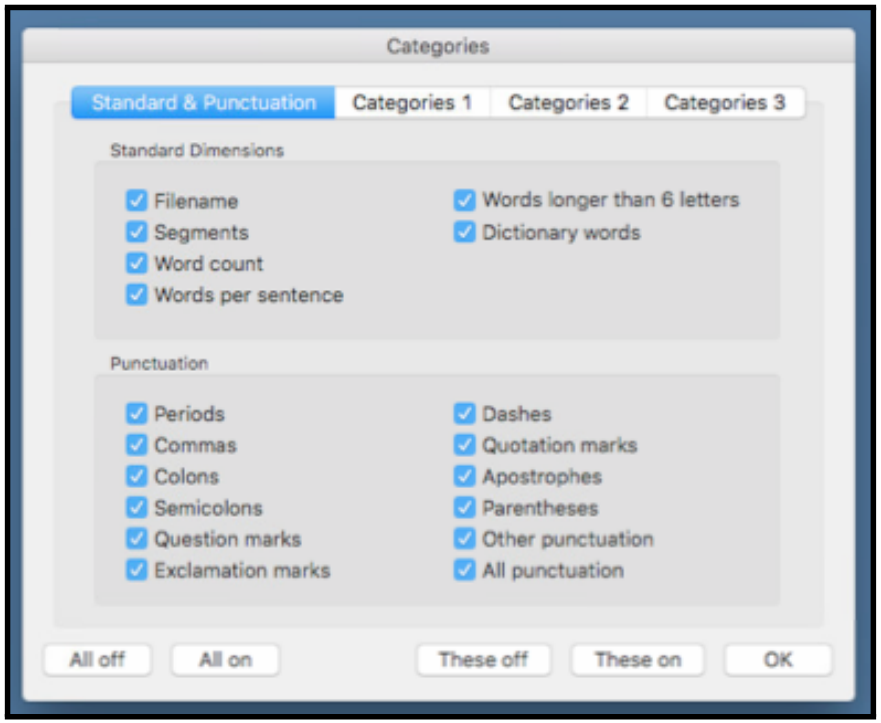
LIWC: Modify A Dictionary File



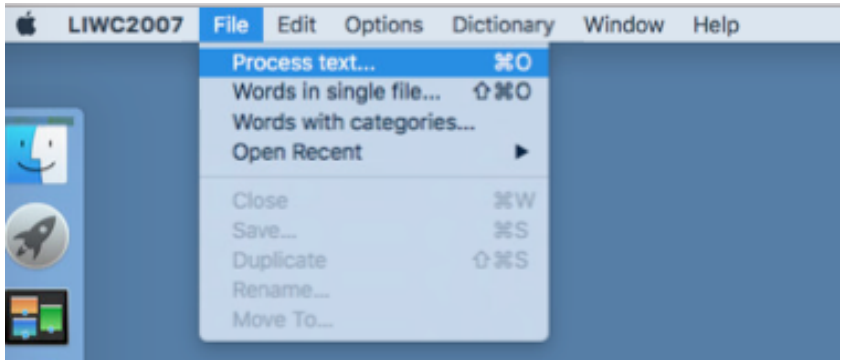


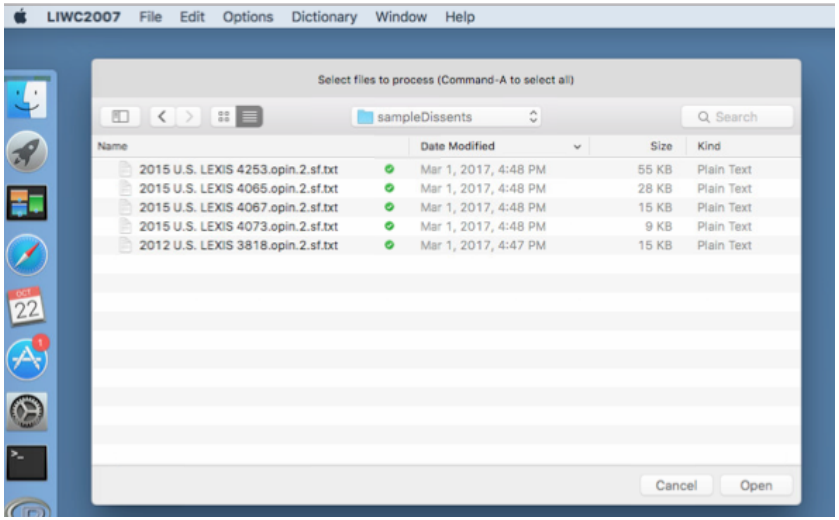
LIWC: Customize Output

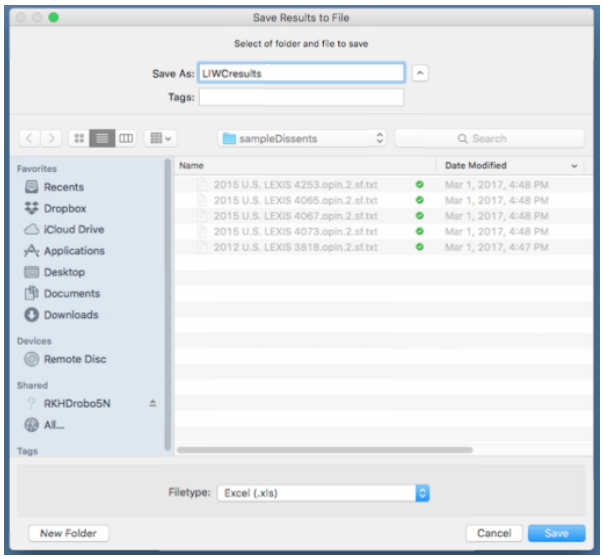




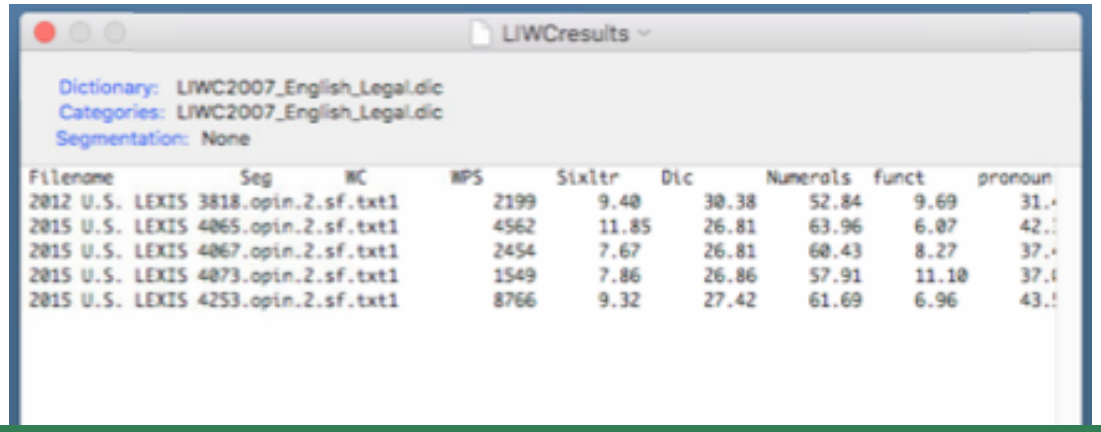
LIWC: Run Analysis







LIWC: Output



LIWCresults

Home | Insert | Page Layout | Formulas | Data | Review | View

Paste | Cut | Copy | Format | Calibri (Body) | 12 | Wrap Text | General | Conditional Formatting | Format as Table

R9

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Filename	Seg	WC	WPS	Sixltr	Dic	posemo	negemo	ppron	past	present	future	preps	shehe
2	2012 U.S. LEXIS 3818.opin.2.sf.txt	1	2199	9.4	30.38	52.84	0.82	0.23	0.41	1.64	3.23	0.86	13.01	0.05
3	2015 U.S. LEXIS 4065.opin.2.sf.txt	1	4562	11.85	26.81	63.96	0.9	0.75	2.08	0.81	4.08	1.36	12.87	0.28
4	2015 U.S. LEXIS 4067.opin.2.sf.txt	1	2454	7.67	26.81	60.43	2.36	0.94	1.55	1.26	4.69	0.9	10.64	0.04
5	2015 U.S. LEXIS 4073.opin.2.sf.txt	1	1549	7.86	26.86	57.91	2.58	2.13	2.32	1.36	3.74	0.45	11.68	0.77
6	2015 U.S. LEXIS 4253.opin.2.sf.txt	1	8766	9.32	27.42	61.69	1.28	0.7	1.35	1.44	3.4	0.87	12.96	0.13
7														

WordNet: The Basics

- WordNet is a lexical database founded by George A. Miller at Princeton, and is still run by a team at Princeton.
- It is essentially a combination of a dictionary and a thesaurus on steroids.
- It contains over 155,000 words - nouns, verbs, adjectives, and adverbs.
- It does **not** contain determiners, prepositions, pronouns, conjunctions, or particles.
- Homepage: `http://wordnet.princeton.edu`.

WordNet: An Overview of the Nitty Gritty

- All words are organized into synsets with their synonyms. A word with more than one meaning (about 17%) has more than one synset.
- Each synset is linked to other synsets though a variety of lexical relationships. These include:
  - Antonymy - opposite (e.g., wet, dry)
  - Hyponymy - subordinate noun (e.g., tree, plant) Since a tree is a specific kind of plant it is linguistically subordinate.
  - Troponymy - subordinate verb (e.g., march, walk) Since a march is a specific manner of walking it is linguistically subordinate.
  - Meronymy - part (e.g., branch, tree) There is a distinction between component parts, substantive parts, and member parts.

Why use WordNet?

- Identify possible parts of speech for each word
- Measure breadth/specificity of language
- Account for use of synonyms
- Expand the usefulness of an existing dictionary
- Efficiently build a comprehensive dictionary from scratch
- Exclude gobbledygook
- Evaluate OCR (Optical Character Recognition).

Evaluating OCR

No. 94-859

3n tfjc Supreme Court ot tfjc ®mtcb States

October Term, 1994

BRUCE BABBITT, SECRETARY OF THE INTERIOR, ET AL., Petitioners

v.

SWEET HOME CHAPTER OF COMMUNITIES FOR A GREAT OREGON, ET AL., RESPONDENTS

On Writ of Certiorari to the United States Court of Appeals for the District of Columbia Circuit

BRIEF AMICUS CURIAE OF THE AMERICAN FARM BUREAU FEDERATION, CALIFORNIA FARM BUREAU FEDERATION, OREGON FARM BUREAU FEDERATION, AND TEXAS FARM BUREAU IN SUPPORT OF RESPONDENTS*

INTERESTS OF THE AMICI CURIAE

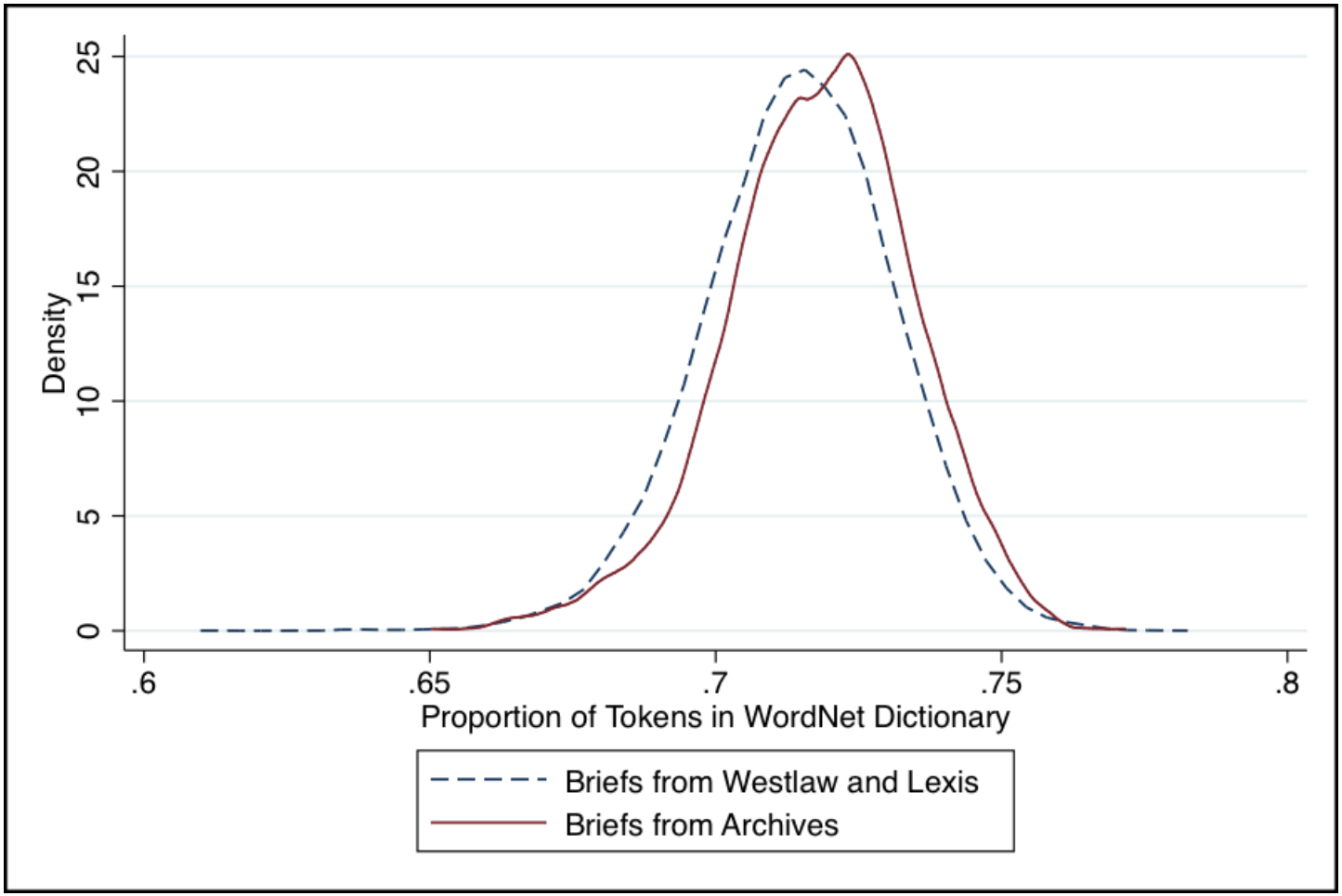
The American Farm Bureau Federation (AFBF) is a voluntary general farm organization organized in 1920 under the General Not-For-Profit Corporation Act of the State of Illinois. AFBF was founded to protect, promote, and represent the business, economic, social and educational interests of American farmers and ranchers. AFBF has member organizations in all 50 states and Puerto Rico

Consents to the filing of this brief are on file with the Clerk.

2

representing more than 4.4 million member families. Amici California, Oregon, and Texas Farm Bureaus are members of AFBF, representing the interests of farmers and ranchers in their respective states.

Evaluating OCR Performance with WordNet



Installing Python (and other goodies)

- Install Anaconda (www.anaconda.com)
- Now you have Python, Jupyter, and NLTK
- Jupyter Notebook is a handy way to run Python code
- Go to Terminal/Command Line and type:
  `jupyter notebook`

Pre-processing Data

- Taking out the trash: What do we want to ditch?
  - citations
  - page numbers
  - non-words
  - gobbledy-gook
  - stopwords
- Making life easier: Boiling things down
  - lemmatizing
  - stemming
- Summarizing Text
  - Frequency distributions
  - Plots and summary statistics