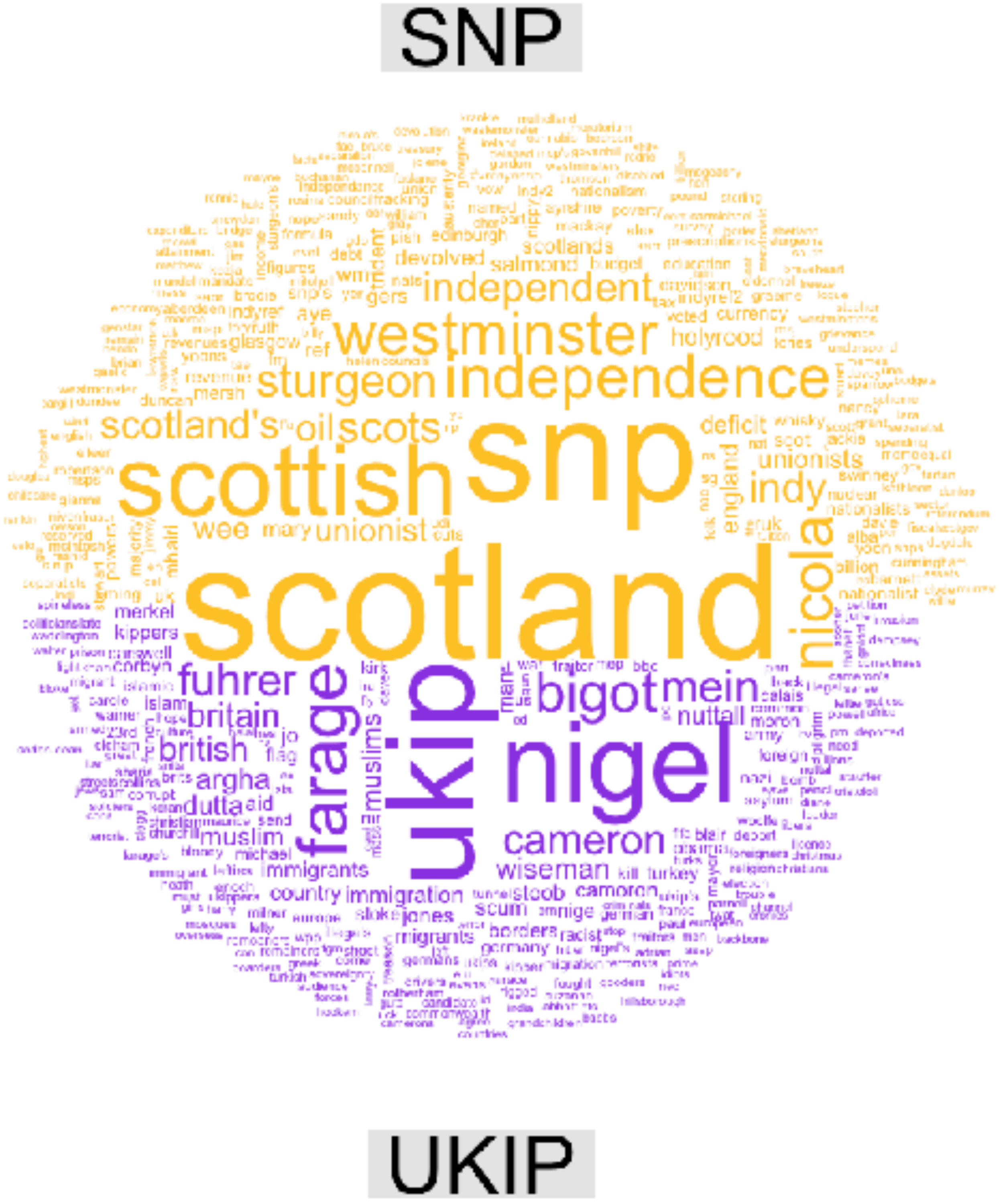
# Basic Text Analysis using R

Justin Ho

# What is Text Analysis?



“Systematic, objective, quantitative analysis of message characteristics"

*–Neuendorf, The Content Analysis Guidebook*

# Different Types of Text Analysis

- Degree of human involvement
  - Human coding (100%)
  - Supervised
  - Unsupervised (0%)
- Type of output
  - Scaling
  - Classification

# 4 Principles of Text Analysis

1. All Quantitative Models of Language Are Wrong
   —But Some Are Useful

2. Quantitative Methods Augment Humans, Not Replace Them

3. There Is No Globally Best Method for Automated Text Analysis

4. Validate, Validate, Validate

Read: Grimmer, J. and Steward, B. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267-297. doi:10.1093/pan/mps028

# Bag of Words Assumption

- Word order doesn’t matter
- The following three lines mean exactly the same:
  - I enjoy eating food and being with my family
  - I enjoy eating my family and being with food
  - and being eating enjoy family food I my with

# End Product

## Document-Feature Matrix

| | word*1* | word*2* | … | word*N* |
|---|---|---|---|---|
| doc*1* | 0 | 0 | | 0 |
| doc*2* | 0 | 1 | | 0 |
| … | | | | |
| doc*N* | 2 | 0 | | 3 |

## Meta Data

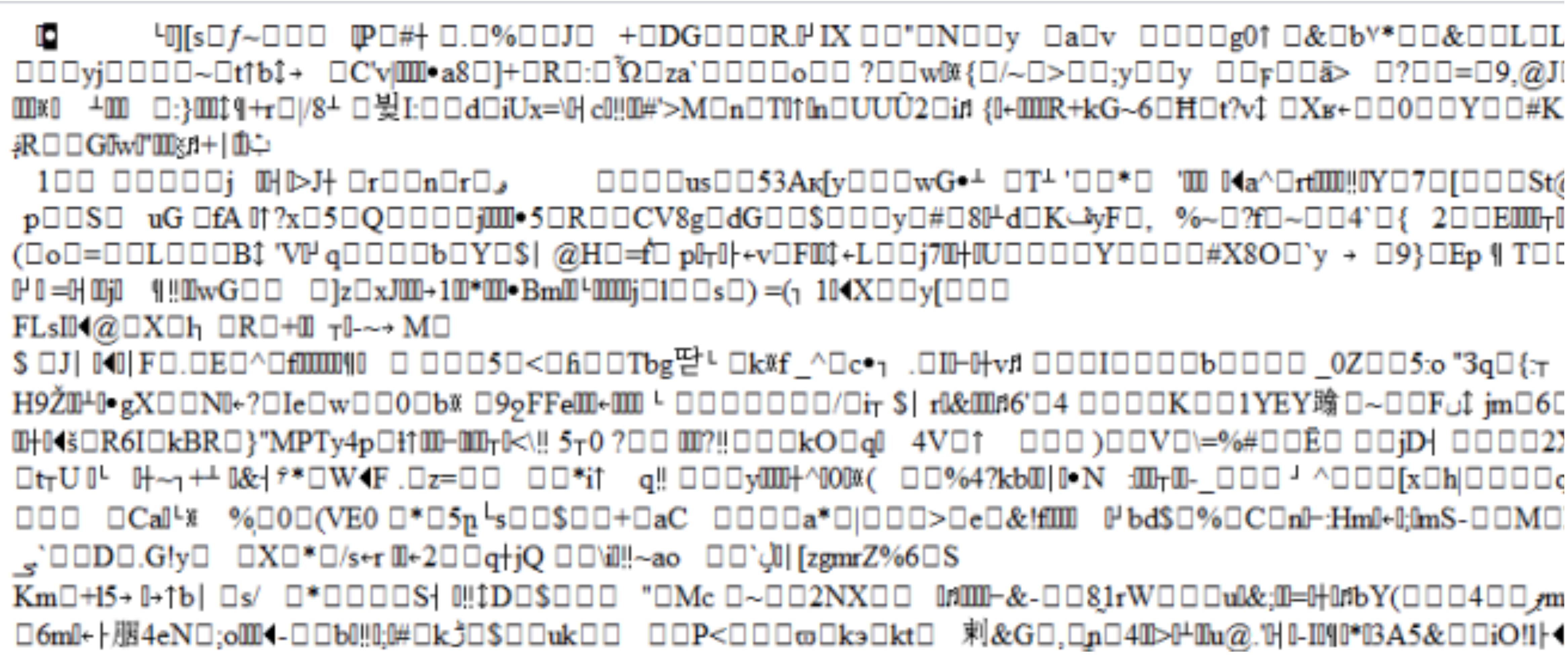
| | var*1* | var*2* | … | var*N* |
|---|---|---|---|---|
| doc*1* | Y | 1 | | 0 |
| doc*2* | N | 1 | | 7 |
| … | | | | |
| doc*N* | Y | 0 | | 10 |

# Workflow of Text Analysis

1. Load the text into R
2. Preprocess
3. Analyse
4. Back to Step 2

# Loading the Text into R: Encoding

- Common Encoding Standards:
  ASCII, Unicode (UTF-8/16)



# Loading the Text into R: Extracting the Text

- Extracting useful Text:
  - <h1 class="chapter0" id="c01"><a id="page13"></a><strong>BRAN</strong></h1><p class="nonindent"><span class="dropcaps">T</span>he morning had dawned clear and cold, with a crispness that hinted at the end of summer. They set forth at daybreak to see a man beheaded, twenty in all, and Bran rode among them, nervous with excitement. This was the first time he had been deemed old enough to go with his lord father and his brothers to see the king’s justice done. It was the ninth year of summer, and the seventh of Bran’s life.</p>

# Preprocessing

- Remove capitalization, numbers, and punctuation
- Before:
  Article 1. All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.
- After:
  article all human beings are born free and equal in dignity and rights they are endowed with reason and conscience and should act towards one another in a spirit of brotherhood

# Preprocessing

- Remove stop words
- Before:
  article all human beings are born free and equal in dignity and rights they are endowed with reason and conscience and should act towards one another in a spirit of brotherhood
- After:
  article all human beings born free equal dignity rights endowed reason conscience act one another spirit brotherhood

# Preprocessing

- Apply stemming algorithm
- Before:
  article all human beings born free equal dignity rights endowed reason conscience act one another spirit brotherhood
- After:
  articl all human be born free equal digniti right endow reason conscienc act one anoth spirit brotherhood

# Preprocessing

- Create Document-Feature Matrix

| | articl | all | human | be | born | free | equal | … | brother hood |
|---|---|---|---|---|---|---|---|---|---|
| Doc1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | … | 1 |

# Time for R…

https://github.com/justinchuntingho

# Keyword Analysis

- What is a Keyword?
  - “A keyword may be defined as a word which occurs with unusual frequency in a given text. This does not mean high frequency but unusual frequency, by comparison with a reference corpus of some kind”

    (Scott, M. (1997). PC analysis of key words - and key key words. *System*, 25(2), 233-45.)

# Keyword Analysis

- What is Keyness?
  - “The keyness of a keyword represents the value of log-likelihood or Chi-square statistics; in other words it provides an indicator of a keyword’s importance as a content descriptor for the appeal.”
- What is a Chi-squared test?
  - Comparison between the observed frequency and expected frequency.

# Keyword Analysis

- What is p-value?
  - “The significance (p-value) represents the probability that this keyness is accidental”.
  - The odds of observing the same observation in a purely random world.

(Biber, D., Connor, U. & Upton, A. with Anthony, M. & Gladkov, K. (2007). Rhetorical appeals in fundraising. In D. Biber, U. Connor & A. Upton. *Discourse on the Move: Using corpus analysis to describe discourse structure.* (121-151). Amsterdam: John Benjamin.)

Time for R…

# Want to know more?

- http://quanteda.io/
- http://www.r-bloggers.com/
- Edinburgh Text Analysis Research Group
  - https://jiscmail.ac.uk/TEXTANALYSIS
  - https://edtextanalysis.wordpress.com/

# My Contact

- Email: Justin.Ho@ed.ac.uk
- Twitter: @zjustin334
- Github: https://github.com/justinchuntingho