Centre for
Data Analytics

Insight

Critical Discourse Analysis + Machine Learning

Igor Brigadir, Derek Greene, Pádraig Cunningham

Analysing Discourse: Theories, Methods & Applications (Mar 4th, 2015)

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN sfi Science Foundation Ireland

Problem:

Why CDA?
Offers understanding beyond Retrieval, Summarization.

CDA is hard! But: Open to new approaches.

Need Objective, Systematic, **Reproducible** Results.

Need to work with Large Corpora: #indyref Data

- Yes: 618 Accounts, 799,096 Tweets
- No : 610 Accounts, 570,024 Tweets
- Timespan: 70 Days, 11th Aug 2014 to 19th Oct 2024

Corpus Assisted Critical Discourse Analysis:

"... social power abuse, dominance, and inequality are enacted, **reproduced...** by **text** ..."
(Van Dijk, 2008)

Can Existing Language Modelling, Natural Language Processing techniques help?

Corpus Linguistics already being applied!
"Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word 'Muslim' in the British Press 1998–2009"
(Baker et al , 2012)

Map CDA theory to NLP Methods: Potentially improving state-of-the-art in both!

Corpus Assisted Critical Discourse Analysis:

Steps:

- Problematization
- Theorization
- Contextualization
- Methodology & Analysis
- Interpretation

Where Text Mining can help:

- Data Gathering, Building Corpora
- Supervised / Unsupervised Learning
- Lexis, Style, Genra Classification at scale.
- Measurement Validity!

Something Useful: 2-fold Cross Validation

CDA “over-analyzed” == ML “overfitting”

Steps:

- Shuffle Documents
- Split in 2 sets
- Perform CDA independently on both (Train)
- Compare Results (Test)
- Quick & Dirty Approach to Measurement Validity.

Methods: Overview

Scottish Independence on Twitter:

- Split By Group & Over Time (7 Day Window)
- Train Distributional Semantic Models (Collocations)
- for each Community & Each “Window”

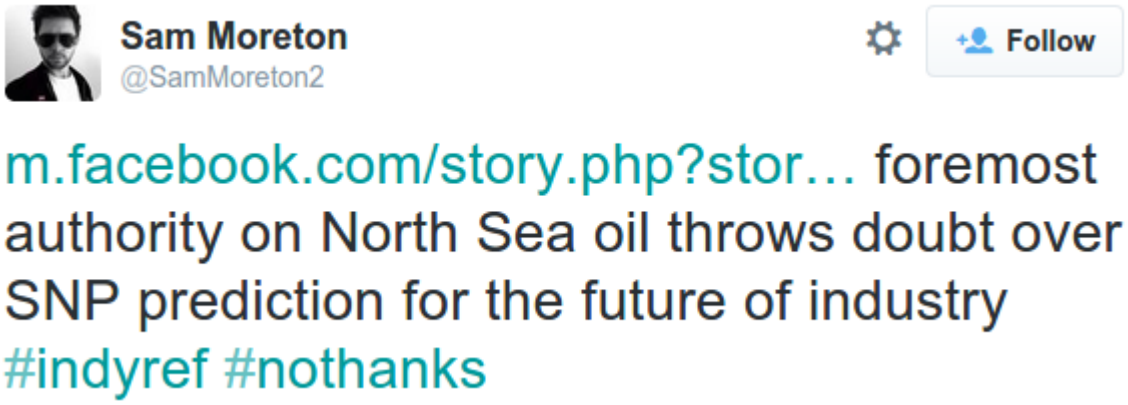
- Overall Similarities? Differences?
- Changes over time within Community
- Changes over time across Communities
- Explore “Unusual” Patterns

word2vec & “keyword-in-context”

Learn a word vector representation that is good at predicting the nearby words.

Given a word, predict the surrounding context

Example:



word2vec & “keyword-in-context”

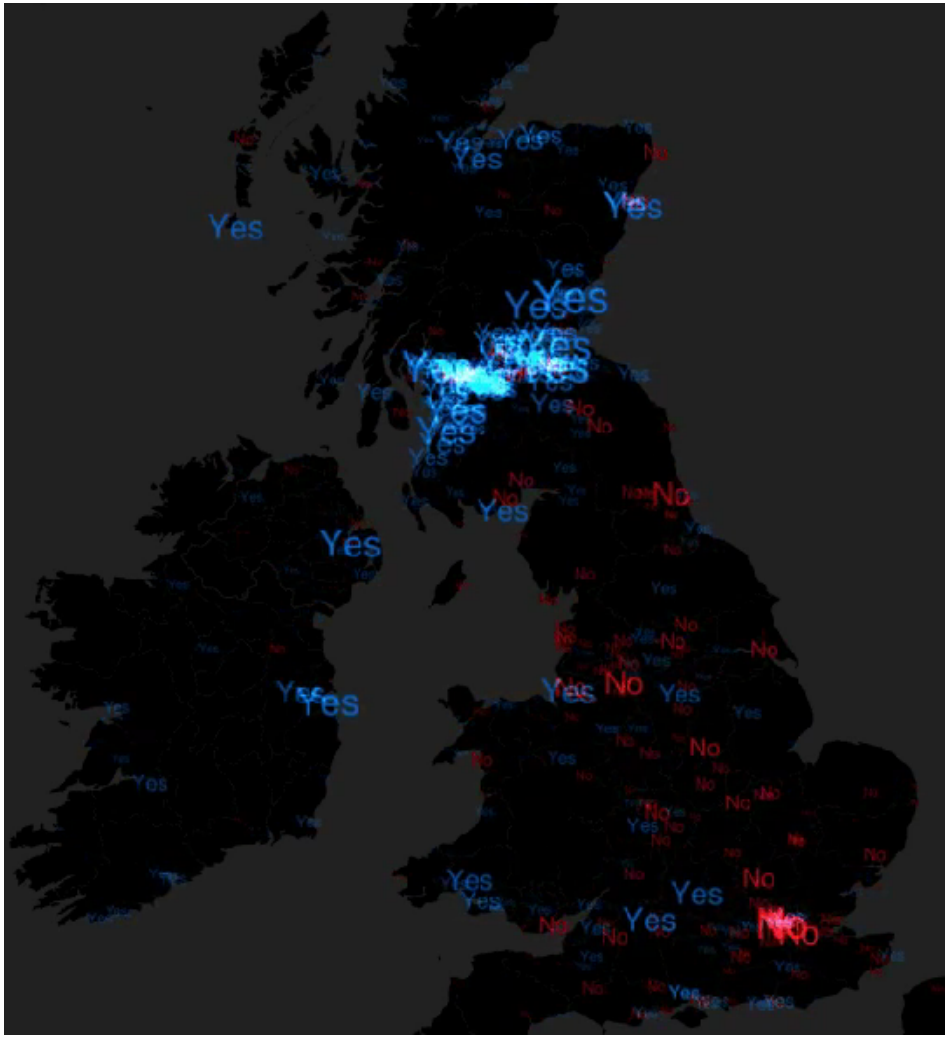
```
yref foremost authority on north sea oil throws doubt over
…
lmonds bizarre blockade of north sea comment @ sygazette #
 oil and gas riches of the north sea are entirely different
…
foremost authority on north sea oil throws doubt over snp p
…
s bizarre blockade of north sea comment @ sygazette # indyr
and gas riches of the north sea are entirely different who
…
most authority on north sea oil throws doubt over snp predi
…
es and wont come back 2 the oil numbers will get @ politica
 tweet poverty in scotlands oil capital my @ dw_english rep
had to speak out to warn of oil depletion # indyref cant re
```

Scottish Independence:

Yes Campaign was MUCH more active on Twitter.

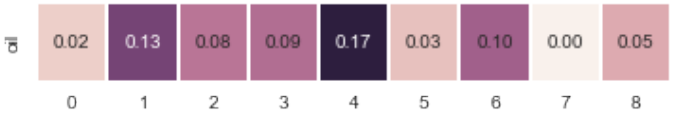
The topic of was North Sea Oil featured heavily.

How Did the Conversation change over time?

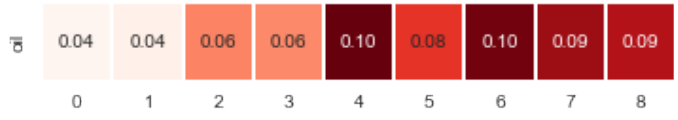


Scottish Independence:

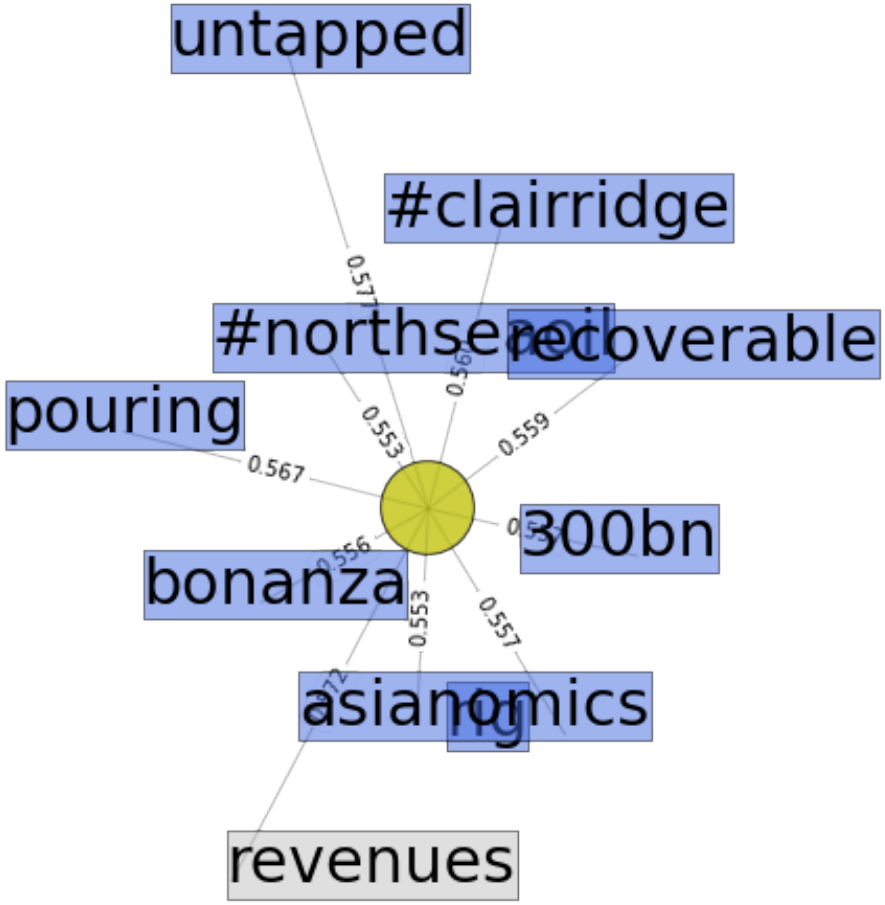
t = 3	t = 4	t = 5	t = 6
untapped	untapped	norways	fields
revenues	yada	revenues	reserves
pouring	reserves	reserves	undercutting
#clairridge	bonanza	147bn	magically
recoverable	@energyvoicenews	discoveries	boosts
300bn	revenues	chevron	#scotbiz
rig	bloddy	bonanza	clair
bonanza	nth	snake	unspecified
#northsea	1billion	amass	production
asianomics	commercially	@jamesclats	andtech
norways	discoveries	offshore	ns

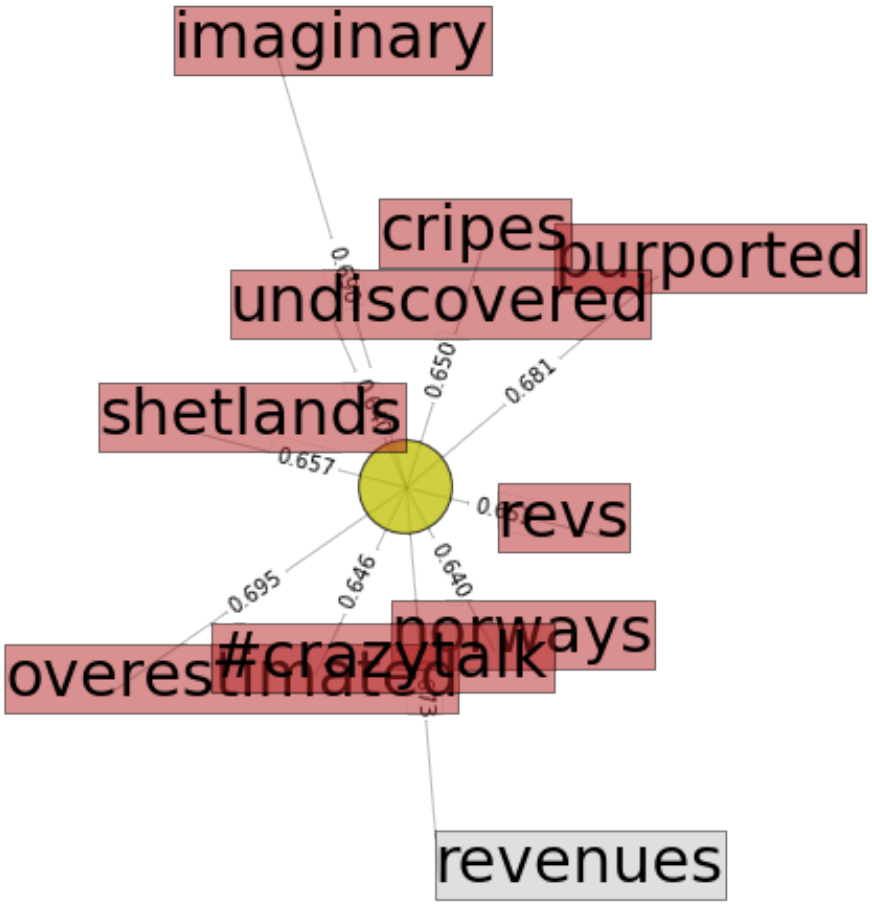


t = 3	t = 4	t = 5	**t = 6**
overestimated	fields	revenues	salesmen
imaginary	rigs	ridge	@davidgjonesss
purported	rosy	bonanza	humongous
revenues	revenues	sea	revenues
shetlands	downgrade	4bn	gigantic
revs	@nu2socmed	offset	pollutes
cripes	bonanza	estimates	depletes
#crazytalk	inflated	snake	snake
norways	generated	geology	1986
undiscovered	@huwgruffydd	yoyo	geographical



Scottish Independence:





Future Work:

Is it actually useful in any way? **Reproduce Existing Work**, or Develop Novel Lines of Inquiry.

More Data:

- US Midterm Elections (Republicans / Democrats)
- Irish Elections: Dáil & Seanad (Irish Politicians on Twitter)
- Pro / Anti Vaccination Communities (Ongoing)

Word Space Models: Better Measures? Visualisation?

References:

Brigadir, I.; Greene, D.; and Cunningham, P. 2015. Analysing Discourse Communities With Distributional Semantic Models. *Forthcoming*

Van Dijk, T. A. 2008. “Critical discourse analysis. The Handbook of Discourse Analysis”.

Porter, J. 1992. “Audience and rhetoric: an archaeological composition of the discourse community”. Prentice Hall.

Fairclough, N. 1995. “Critical Discourse Analysis: Papers in the Critical Study of Language.” Language in social life series. Longman.

Firth, J. 1957. “A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis.” Philological Society, Oxford.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. “Efficient estimation of word representations in vector space. CoRR abs/1301.3781.

Centre for
Data Analytics

Insight

github.com/igorbrigadir/CDA

UCD Science East, 3rd Floor

DCU

NUI Galway
OÉ Gaillimh

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

UCD
DUBLIN

sfi
Science Foundation Ireland