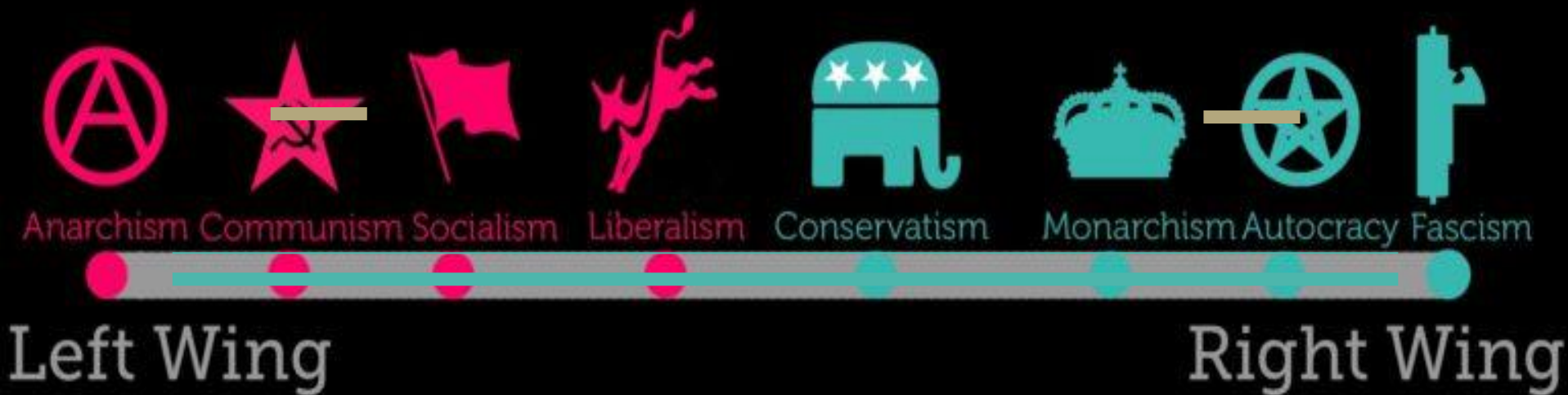# The words are important!

Political Lexicon

From the sophists vs Socrates to Cambridge Analytica, From the speech in the street to the tweet: The art of manipulation of language.

Text classification

Opinion mining and text classification are still very restricted. Text, words and thoughts are dynamic and fluid.
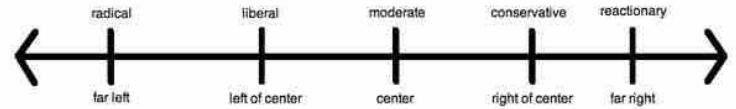
The goal

Detect the ideology of a sentence

Using quotes of people with a strong political bias try to identify and capture key-words and pattern to detect ideology hidden behind

The Dataset

The political spectrum

	Authoritarianism		
Stalin Communist Dictators	Generic	Tyrants	Nazis Fascists
State Socialism	Social Democracy	Democrat Party	Republican Party
Democratic Socialism	Social Liberalism	Conservative Liberalism	Tea Party
Libertarian Socialism	Classical	Liberalism	Right-wing Libertarianism and Minarchism
	Liberalism		

radical — liberal — moderate — conservative — reactionary

far left — left of center — center — right of center — far right

Data: Selection and Criteria

Criteria

- **Research of not politician people with strong political bias**
- **English mother tongue**
- **Wikipedia confirmation of the bias**
- **Allsides.com**
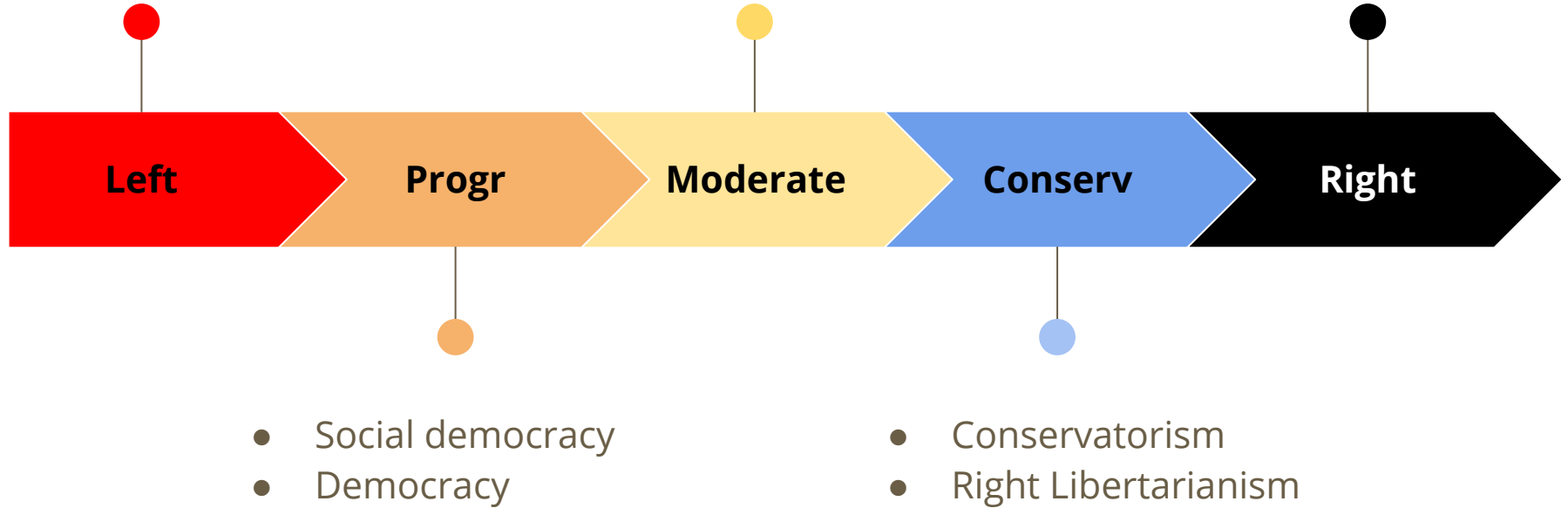- **Not get involved**

Selection

- **Wikiquote**
- **Goodreads**
- **BrainQuote**
- **Random sample to balance classes**

The target

- Communism
- Socialism
- Ambientalism
- Left Anarchism

- Indipendent
- Popularism
- Christian democracy
- Centre Libertarianism

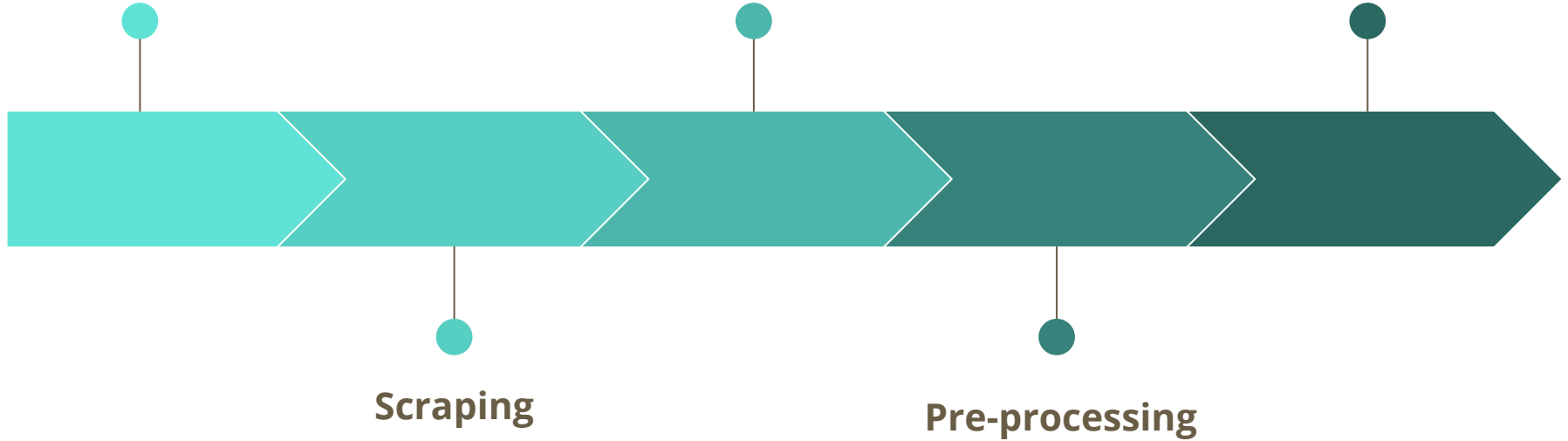- Fascism
- Alt-Right

| Left | Progr | Moderate | Conserv | Right |

- Social democracy
- Democracy

- Conservatorism
- Right Libertarianism

First steps

Research

Quote in the quote

Lemmatization

Scraping

Pre-processing

Balance

Selection

Baseline

- The weakest category choose the number of the sample
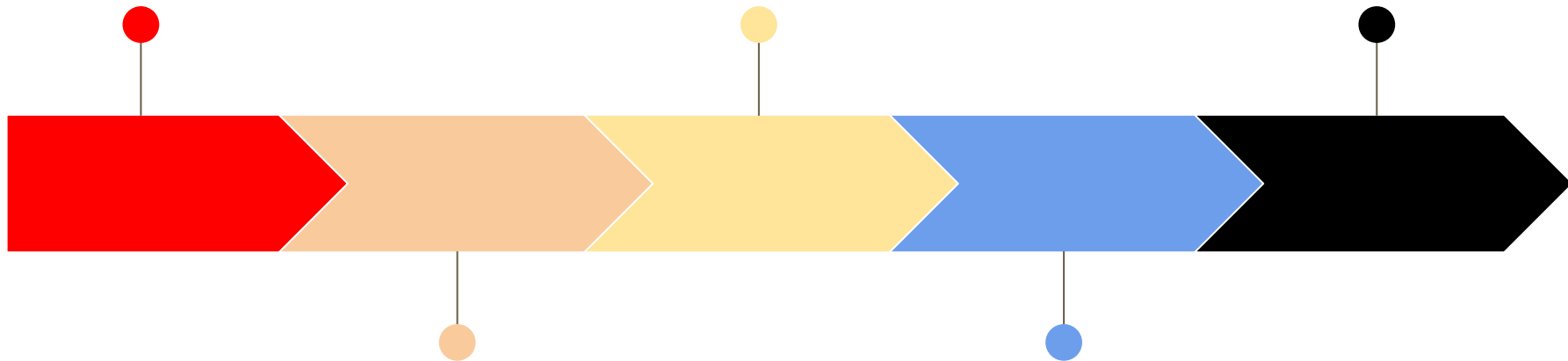- External Dataset with four new authors for each category

0.2

The Vocabulary

- 500 000 total words (after stopwords)
- 27000 unique
- Left is the most chatty: 115000 words
- Moderate the quiet one: 83000
- Longest quote 761 words from Right
- 140 Authors in total
- Right most represented

- 115000 words
- 16000 unique

- 91000 words
- 10000 unique

- 83000 words
- 10000 unique

- 92000 words
- 13000 unique

- 106000 words
- 13000 unique

Confusion will be my epitaph

	Left	Progressive	Moderate	Conservative	Right
0	(people, 368)	(people, 302)	(people, 404)	(people, 355)	(people, 473)
1	(world, 271)	(state, 191)	(world, 230)	(government, 212)	(state, 226)
2	(human, 248)	(world, 170)	(make, 184)	(american, 151)	(world, 211)
3	(state, 234)	(time, 169)	(thing, 163)	(right, 151)	(american, 193)
4	(life, 216)	(government, 160)	(life, 157)	(life, 148)	(right, 186)
5	(time, 213)	(right, 141)	(way, 153)	(state, 146)	(america, 185)
6	(way, 200)	(life, 140)	(moral, 148)	(make, 141)	(muslim, 176)
7	(war, 193)	(way, 128)	(time, 146)	(time, 139)	(white, 164)
8	(right, 182)	(thing, 127)	(good, 145)	(know, 138)	(government, 159)
9	(new, 177)	(know, 127)	(human, 128)	(thing, 135)	(black, 156)
10	(power, 156)	(country, 127)	(think, 128)	(year, 129)	(know, 155)

Modelling

- **CountVectorizer**
  - LogisticRegression
  - MultinomialNB

- **TfiVectorizer**
  - LogisticRegression
  - MultinomialNB

Metrics

- Score not too important
- Precision / Recall
- Probability and threshold
- Difference in the 2 Dataset
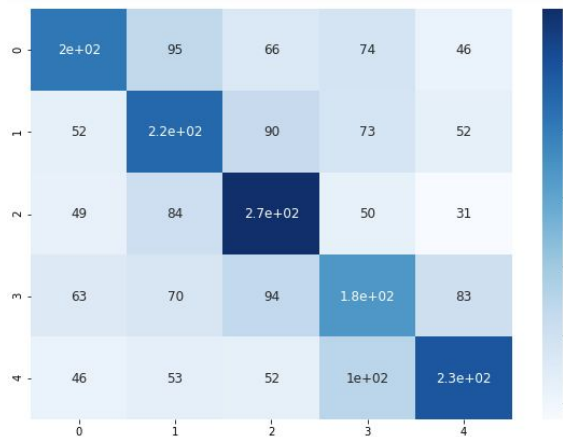- Checking manually

Score
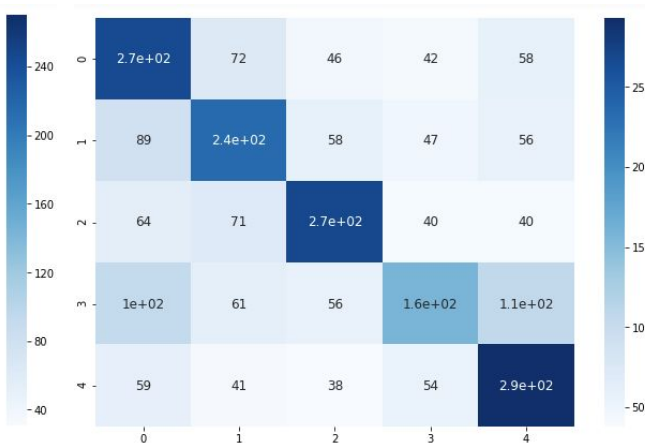
Train / Test

External Dataset

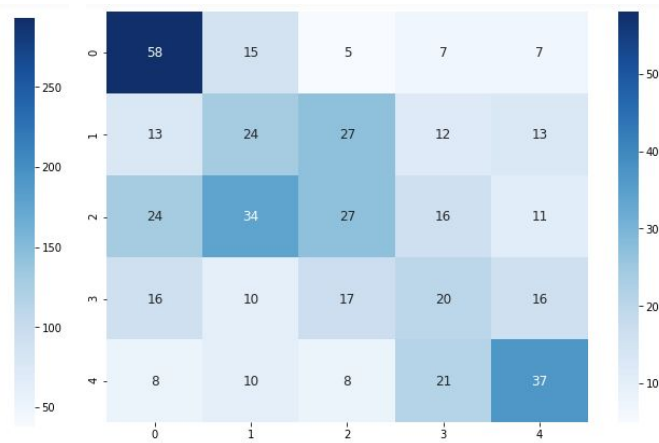0.45 - 0.52

0.28 - 0.32

Confusion will be my epitaph pt. 2

Logistic Regression Train-Test

MultinomialNB Train-Test

External DF

Confusion will be my epitaph pt. 2

MultinomialNB
Train-Test

	precision	recall	f1-score	support
0	0.45	0.55	0.50	485
1	0.50	0.48	0.49	485
2	0.58	0.54	0.56	484
3	0.45	0.33	0.38	485
4	0.52	0.61	0.56	485
micro avg	0.50	0.50	0.50	2424
macro avg	0.50	0.50	0.50	2424
weighted avg	0.50	0.50	0.50	2424

External DF

	precision	recall	f1-score	support
0	0.49	0.63	0.55	92
1	0.26	0.27	0.26	89
2	0.32	0.24	0.28	112
3	0.26	0.25	0.26	79
4	0.44	0.44	0.44	84
micro avg	0.36	0.36	0.36	456
macro avg	0.35	0.37	0.36	456
weighted avg	0.35	0.36	0.36	456

Words for Left

LR -- CV	NB -- CV	LR -- Tfi	NB -- Tfi

POS

- **Climate**
- **Capitalism**
- **Corporate**
- **Class**

NEG

- America
- Muslim

POS

- **People 0.21**
- **World 0.18**
- **Human 0.16**
- **State 0.14**
- **Life 0.14**

POS

- **Human**
- **Climate**
- **Woman**
- **Capitalism**

NEG

- America
- Moral

POS

- **Human 0.0027**
- **Woman 0.0025**
- **War 0.002**
- **Climate 0.002**
- **Class 0.0018**

Words for Progressive

| LR -- CV | NB -- CV | LR -- Tfi | NB -- Tfi |

POS

- City
- NSA
- Open Society
- Help

NEG

- Evil
- God

POS

- People 0.19
- State 0.12
- World 0.11
- Time 0.1
- Government 0.1

POS

- City
- Economic
- NSA
- Copyright

NEG

- Black
- Human

POS

- City 0.003
- Book 0.0025
- Market 0.0023
- Culture 0.0023
- Street 0.0021

Words for Moderate

LR -- CV   NB -- CV

POS

- **Company**
- **Fairness**
- **Google**
- **Management**

NEG

- **Black**
- **Civilization**

- **People 0.26**
- **World 0.14**
- **Make 0.13**
- **Life 0.12**
- **Way 0.1**

LR -- Tfi   NB -- Tfi

POS

- **Company**
- **Moral**
- **Work**
- **Group**

NEG

- **Black**
- **White**

- **Company 0.004**
- **Work 0.002**
- **Love 0.002**
- **Need 0.002**
- **Fear 0.0019**

Words for Conservative

LR -- CV | NB -- CV

POS

- Darwinism
- God
- Polygamist
- Public school

NEG

- Company
- Lose

- People 0.23
- Government 0.12
- Life 0.1
- American 0.09
- Right 0.08

LR -- Tfi | NB -- Tfi

POS

- Government
- Freedom
- God
- School

NEG

- Company
- World

- Government 0.003
- Freedom 0.003
- War 0.0028
- Trump 0.002
- Class 0.0018

Words for Right

LR -- CV | NB -- CV

POS

- **Muslim**
- **Hitler**
- **Immigration**
- **Progressive**

NEG

- **Place**
- **Class**

POS

- **People 0.29**
- **State 0.14**
- **World 0.13**
- **America 0.12**
- **Right 0.11**

LR -- Tfi | NB -- Tfi

POS

- **Muslim**
- **Progressive**
- **White**
- **America**

NEG

- **Work**
- **Kind**

POS

- **Muslim 0.004**
- **Black 0.0038**
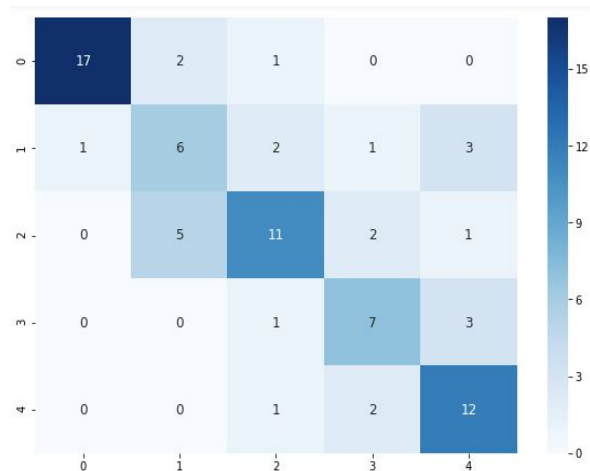- **White 0.0035**
- **Democrat 0.0031**
- **Child 0.0024**

Confusion will be my epitaph pt. 3

Threshold Matrix
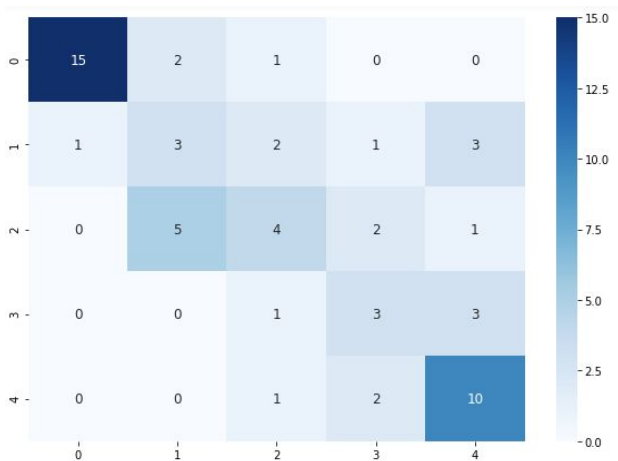
Smooth + Threshold

- Confusion Matrix with Threshold on the probability

- Same Matrix considering the neighbor category

Extra steps

Vader	Word2Vec	N.E.R.

- **Sentiment Analysis = Progressive**

- **Embedding and Parsing = Unclear confusion**

- **Name Entity Recognition = Loss of important features**

"

We don't have to engage in grand, heroic actions to participate in the process of change. Small acts, when multiplied by millions of people, can transform the world.

"

Man vs. Machine

Left

"
We don't have to engage in grand, heroic actions to participate in the process of change. Small acts, when multiplied by millions of people, can transform the world.
"

Howard Zinn

1. [0.44, 0.18, 0.17, 0.13, 0.06]

2. [0.27, 0.18, 0.22, 0.16, 0.15]

3. [0.9 , 0.003, 0.001, 0.002, 0.09]

4. [0.36, 0.19, 0.14 , 0.14, 0.15]

____

Example

"

Obama offered Muslims a careful admonition about women's rights, noting how denying women education impoverishes a country — balanced, of course, with this: "Issues of women's equality are by no means simply an issue for Islam." Example? "The struggle for women's equality continues in many aspects of American life.....

"

<u>Charles Krauthammer</u>

ACTUAL = CONSERVATIVE

1. [0.006, 0.02, 0.1, 0.65, 0.2] *(L)

2. [0.69, 0.001, 0.06, 0.03, 0.21]**

3. [0.25, 0.1, 0.14, 0.17, 0.3]

4. [0.0, 0.0, 0.0, 0.99]

5. [0.19, 0.13, 0.13, 0.19, 0.33]

____

Transformation

"
Public

"
PROGRESSIVE

"
Public school

"
CONSERVATIVE

Man vs. Machine

" Capitalism does a number of things very well: it helps create an entrepreneurial spirit; it gets people motivated to come up with new ideas, and that's a good thing "

Bernie Sanders

PREDICTION

" Conservative , Moderate, Moderate, Moderate "

Future

- **Better Dataset**
- **Different Dataset**
- **Implement an ordinal categorical classification**
- **Investigate about Sentiment Analysis**
- **Topic as feature**
- **Locate space and time**
- **Get a better computer**

Conclusion

- **Homologation of the words**
- **There is a light pattern**
- **A biased person is not always biased**
- **Limitation of ambiguity**
- **The 3 categories in the centre are confused**
- **I have a computational power problem**

Special Thanks to

- My heroic, noisy, slow, dirty, old mac

- And to friends who advised me to change project...They were probably right !