Konzervativec nebo liberál?

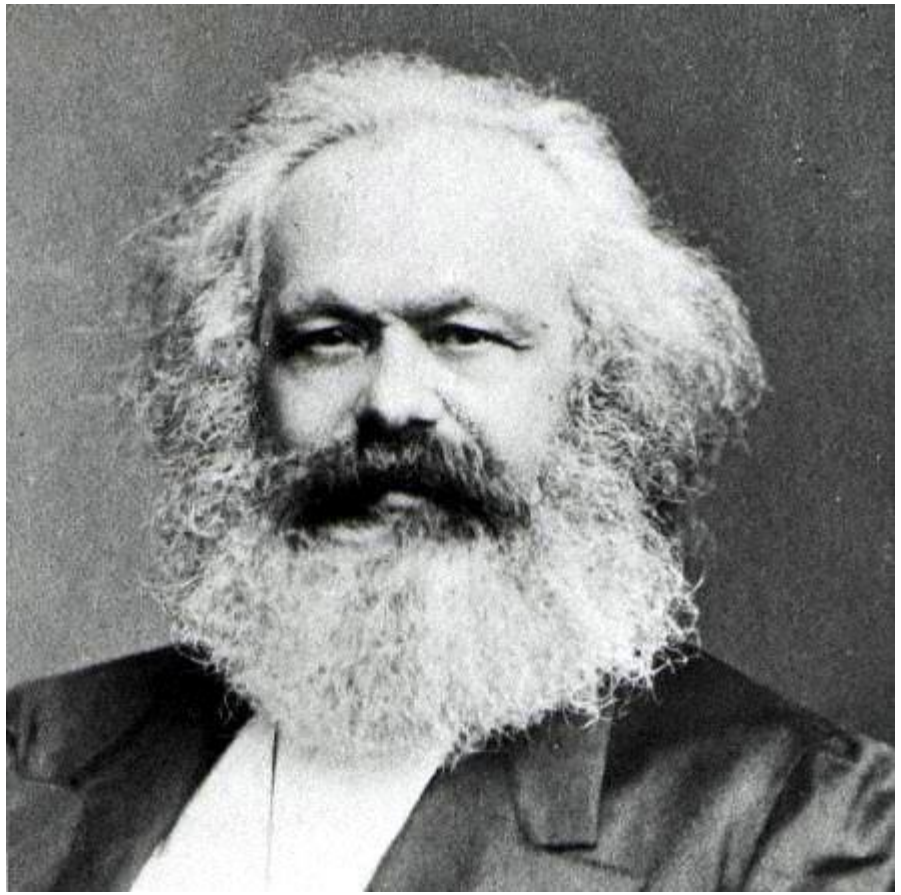
Detekce politické ideologie s užitím rekurzivních neuronových sítí

Vojtěch Škvařil, 399486

PLIN037 Sémantika a počítače, podzim 2014

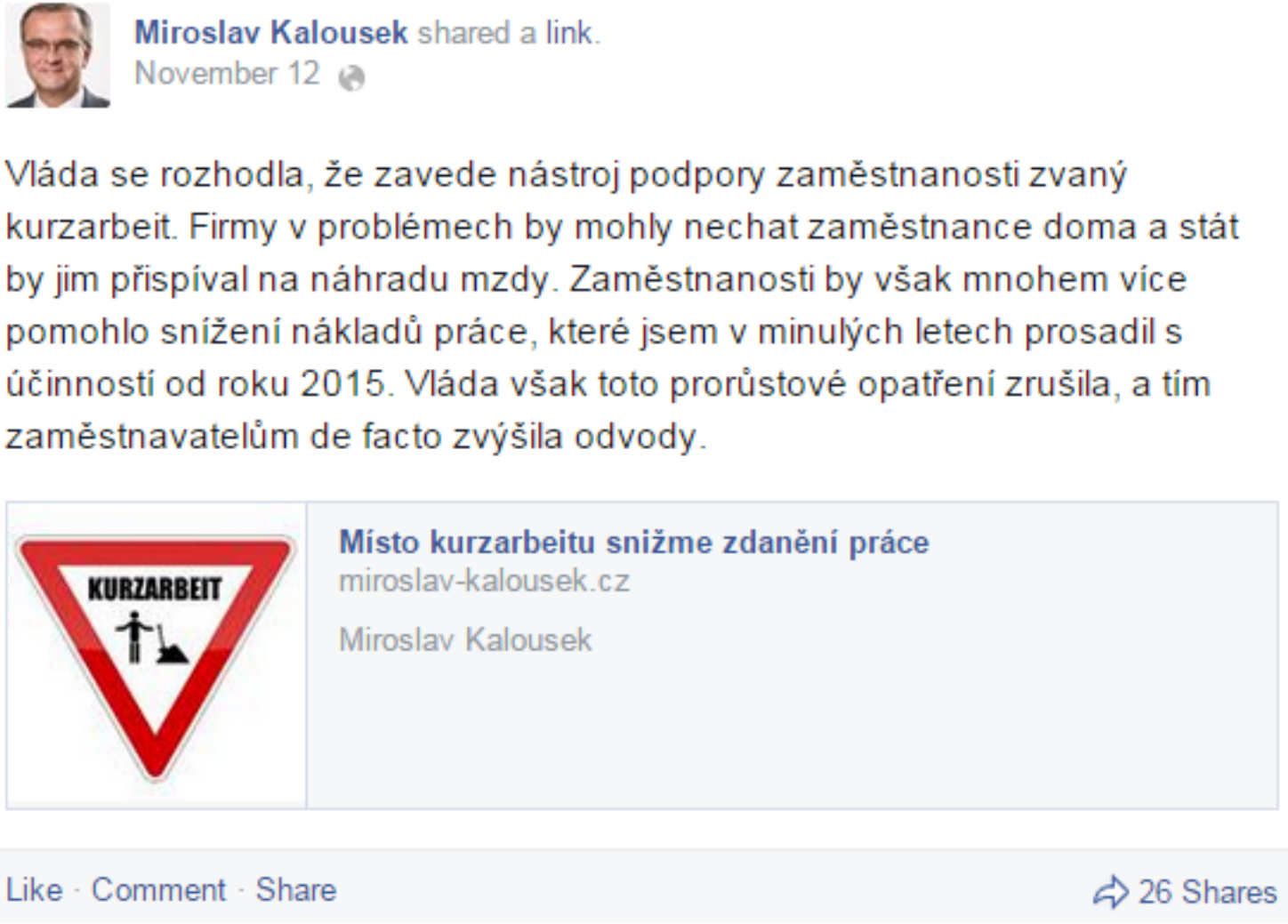
Ideologie

- soustava názorů a hodnot založená na formulování politických, hospodářských a jiných zájmů určité skupiny
- Většinou se jako ideologie označují pouze politické ideologie
- Liberalismus, marxismus, sociální demokracie, …



Politická ideologie v jazykových projevech

- Výběr slov a stavba věty může prozradit jakou politiku pisatel prosazuje
- V USA: pozemková daň X daň za smrt → nic „mezi“

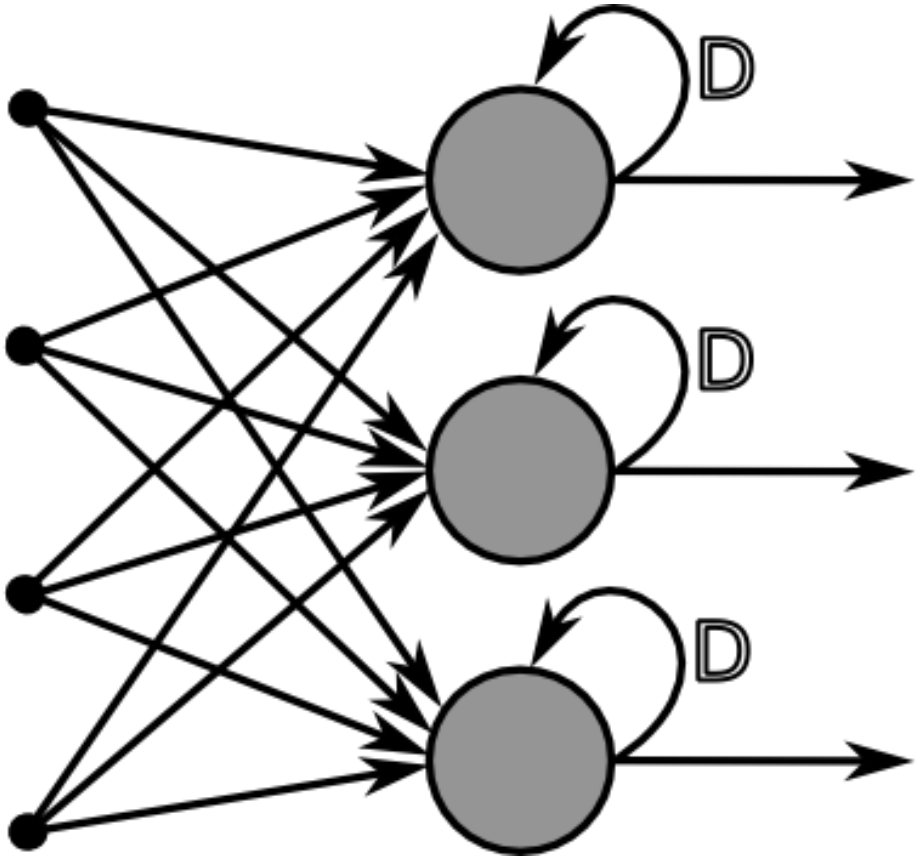


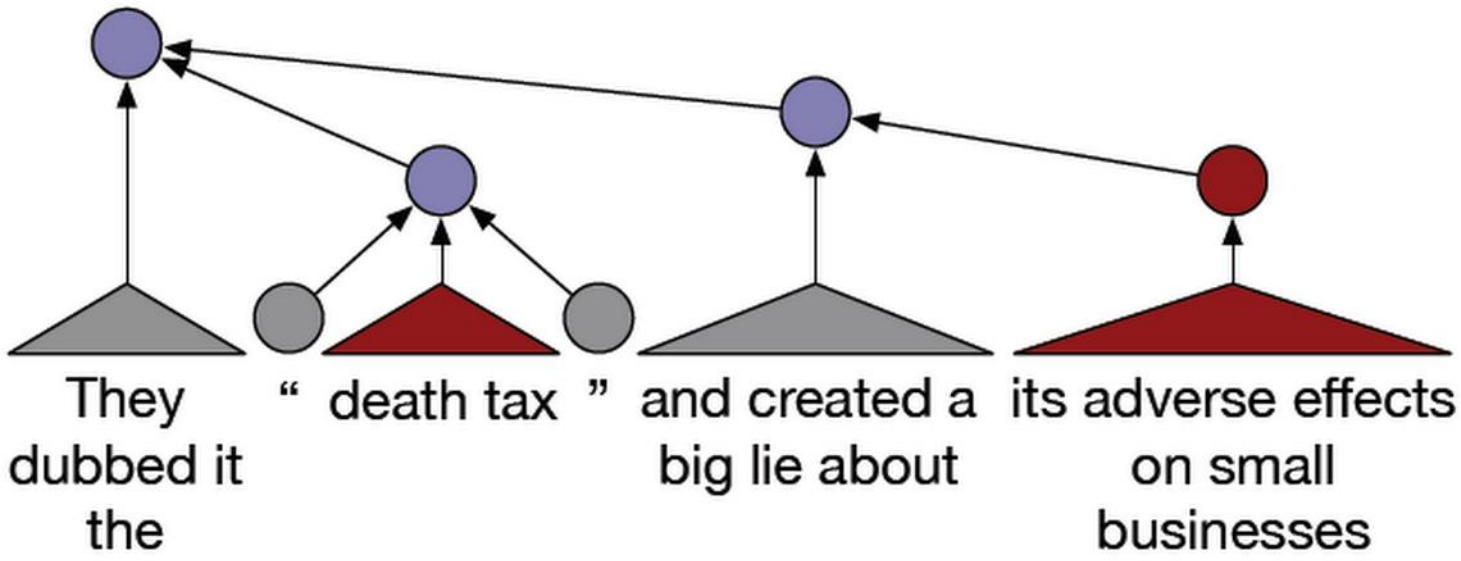
Techniky automatické detekce ideologie

- Bag of words – seznamy slov a jejich významů v daném kontextu
- Sémantické značky na úrovni celého textu nebo celých vět
- Low frequency words – slova, co se vymykají průměrům
- N-gramy
- Implicit sentiment – přesné odvození vztahu slov od struktury věty; ruční anotace, omezený rozsah

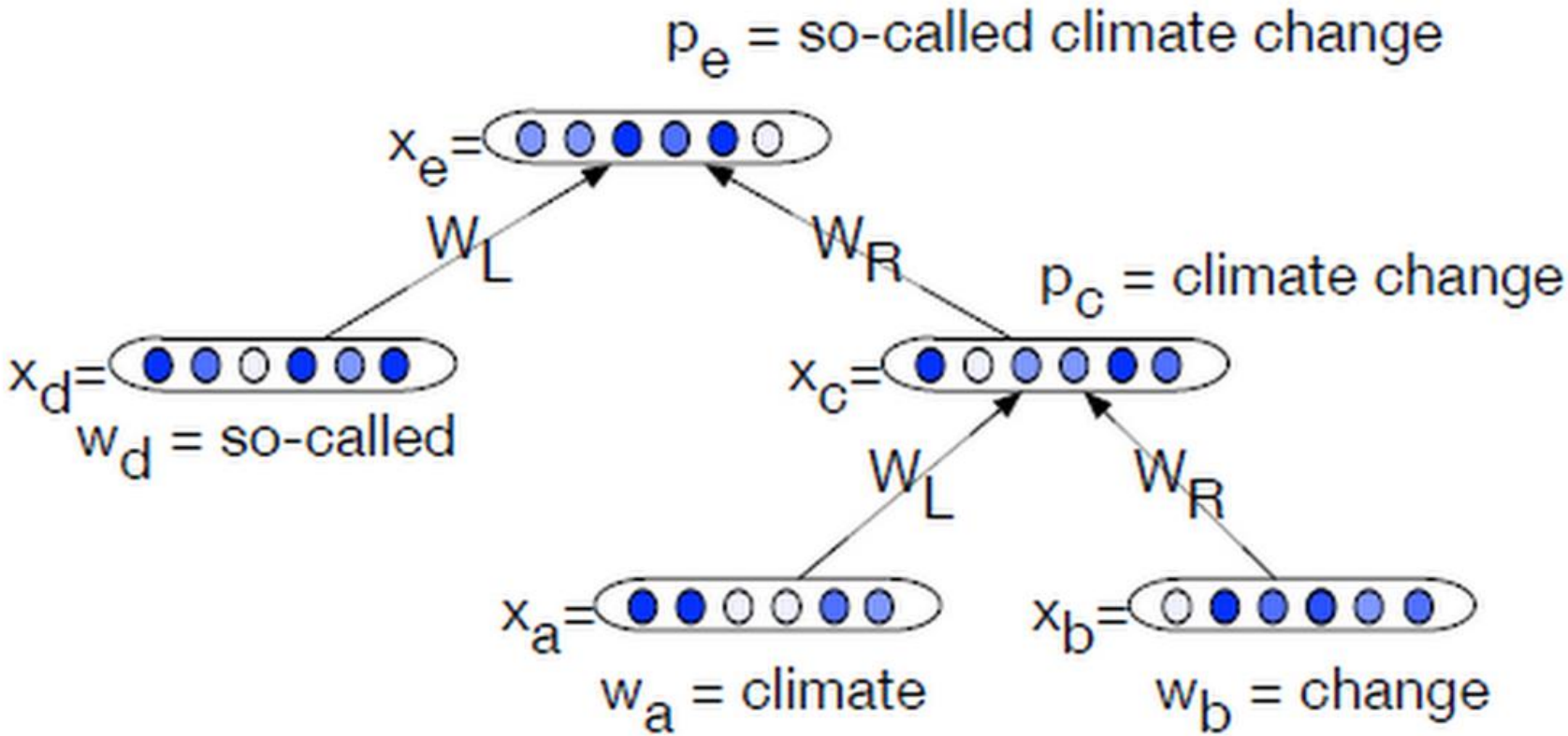
Nový přístup – Rekurzivní neuronové sítě

- RNN – metoda strojového učení, která zachycuje sémantickou a syntaktickou strukturu
- Výborné výsledky u různých detekcí postojů a značkování
- Výhoda RNN → nezávisí na ručně dělaných slovnících, databázích a seznamech pravidel
- Stačí anotované fráze a věty





- RNN mohou modelovat sémantickou kompozicionalitu – ta není implementována univerzálně (ironie, víceznačnost), ale ve většině případů platí
- slova jsou tu reprezentována vektory


p_e = so-called climate change
x_e=
W_L
W_R
p_c = climate change
x_d=
x_c=
w_d = so-called
W_L
W_R
x_a=
x_b=
w_a = climate
w_b = change

Word2vec

- 300dimenzionální nástroj vycházející ze 100 miliardového Google News korpusu, reflektuje význam podobných slov
- Nejbližší společné vektory slov “green” and “energy” obsahují fráze jako “renewable energy”, “eco-friendly” a “efficient lightbulbs”

- Umožní částečně natrénovat RNN algoritmus

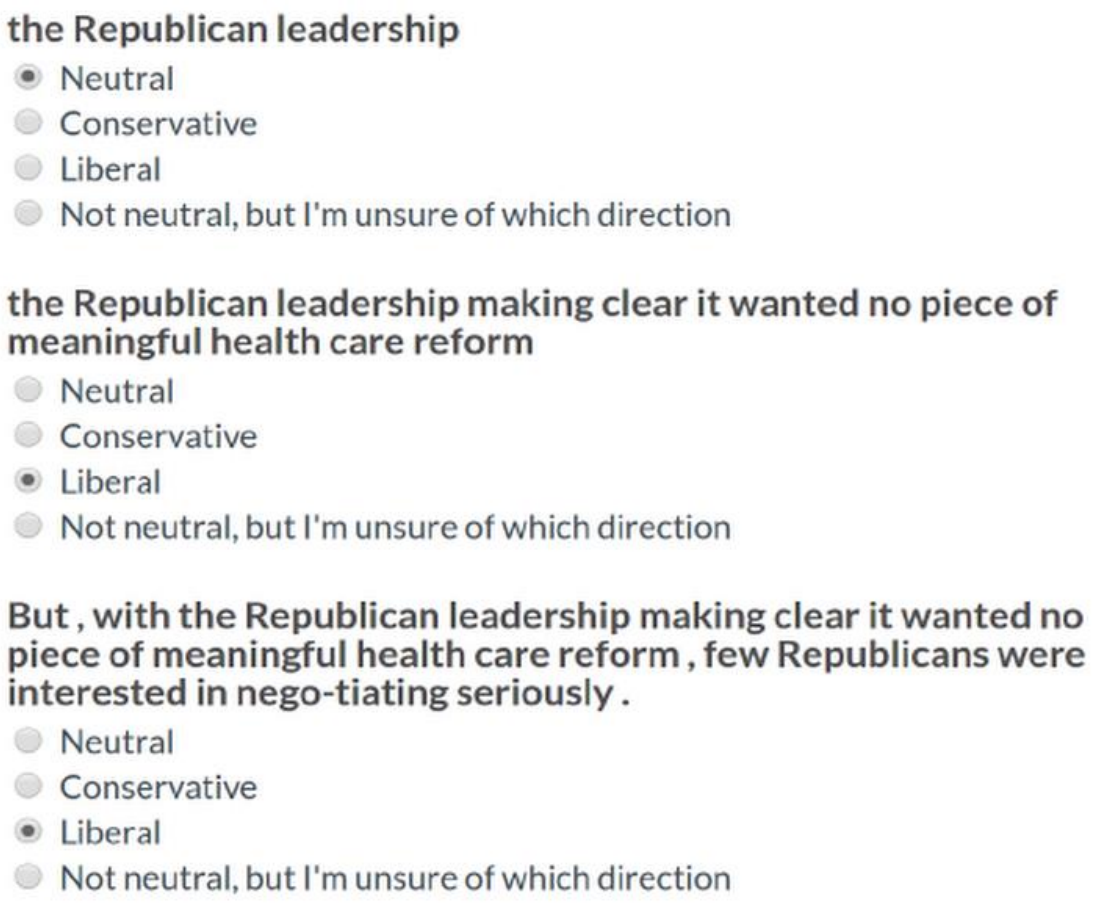
Dataset 1 – projevy v Kapitolu

- Přepisy projevů ve sněmovně → vytvořen dataset značkovaný na úrovni frází a vět
- Které věty jsou dostatečně ideologické? Výběr učiněn přes:
  - LIWC Word Count lexicon rozliší, zda věta obsahuje nějaký názor, určuje 70 různých metrik (pozitivita, styl atp.)
  - “sticky bigrams” – implikují ideologičnost
  - Politici s oblibou používají slova týkající se zabíjení (“masakr”, “porážka”, “poprava”, …)
  - Zbylo tak 7 816 vět

Dataset 2 – Ideological Books Corpus (IBC)

- Kolekce politických textů z knih a článků
- Značkováno na úrovni celého textu, podle světonázoru autora
- Obsahuje i užší vymezení (katolická pravice, libertariánství, …) od politologů
- Milión vět – jak je označkovat? Vybráno 12 tisíc vět.

- Pořadí záměrně od fráze ke větě
- Před vyplněním test způsobilosti
- $0.03 za jednu položku (= 0,7 Kč)
- Každá věta je anotována 3x, stačí shoda 2/3
- Po vyvážení zbylo 3,412 vět



Úspěšnost

- LR = Lexical Represenation
- Použití W2V zvyšuje úspěšnost
- Convote je 2x rozsáhlejší než IBC
- Convote vychází z přepisu řeči, věty jsou kratší a tím i lépe rozpoznatelné
- Neprovedené testy (-) jsou kvůli chybějící vrstvě dat v daném datasetu

Model	Convote	IBC
RANDOM	50%	50%
LR1	64.7%	62.1%
LR2	–	61.9%
LR3	66.9%	62.6%
LR-(W2V)	66.6%	63.7%
RNN1	69.4%	66.2%
RNN1-(W2V)	**70.2%**	67.1%
RNN2-(W2V)	–	**69.3%**

n	Most conservative n-grams	Most liberal n-grams
1	Salt, Mexico, housework, speculated, consensus, lawyer, pharmaceuticals, ruthless, deadly, Clinton, redistribution	rich, antipsychotic, malaria, biodiversity, richest, gene, pesticides, desertification, Net, wealthiest, labor, fertilizer, nuclear, HIV
3	prize individual liberty, original liberal idiots, stock market crash, God gives freedom, federal government interference, federal oppression nullification, respect individual liberty, Tea Party patriots, radical Sunni Islamists, Obama stimulus programs	rich and poor,"corporate greed", super rich pay, carrying the rich, corporate interest groups, young women workers, the very rich, for the rich, by the rich, soaking the rich, getting rich often, great and rich, the working poor, corporate income tax, the poor migrants
5	spending on popular government programs, bailouts and unfunded government promises, North America from external threats, government regulations place on businesses, strong Church of Christ convictions, radical Islamism and other threats	the rich are really rich, effective forms of worker participation, the pensions of the poor, tax cuts for the rich, the ecological services of biodiversity, poor children and pregnant women, vacation time for overtime pay
7	government intervention helped make the Depression Great, by God in His image and likeness, producing wealth instead of stunting capital creation, the traditional American values of limited government, trillions of dollars to overseas oil producers, its troubled assets to federal sugar daddies, Obama and his party as racialist fanatics	African Americans and other disproportionately poor groups; the growing gap between rich and poor; the Bush tax cuts for the rich; public outrage at corporate and societal greed; sexually transmitted diseases , most notably AIDS; organize unions or fight for better conditions, the biggest hope for health care reform

Table 2: Highest probability n-grams for conservative and liberal ideologies, as predicted by the **RNN2-(W2V)** model.

Co s tím?

- pravděpodobnost schválení zákona podle jeho znění
- evoluce rétoriky během předvolební kampaně, od mírné k závěrečné vyhraněné
- “slant index” – jestli se vybrané noviny kloní na určitou ideologickou stranu
- Průzkumy mínění na Facebooku a Twitteru

Děkuji za pozornost

Zdroje

- http://acl2014.org/acl2014/P14-1/pdf/P14-1105.pdf
- Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, Philip Resnik

- https://cs.wikipedia.org/wiki/Ideologie
- http://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf