



UNIVERSITAS
GADJAH MADA

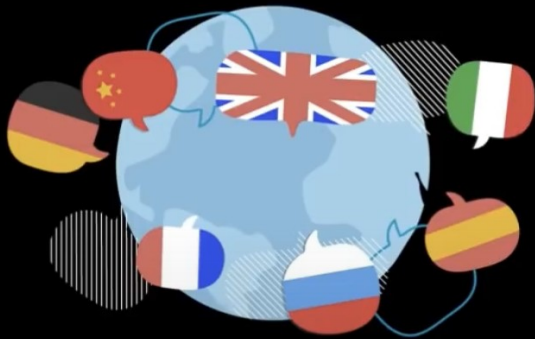


Natural Language Processing

Introduction, theory and application

Faizah, M.Kom

The 21st Century

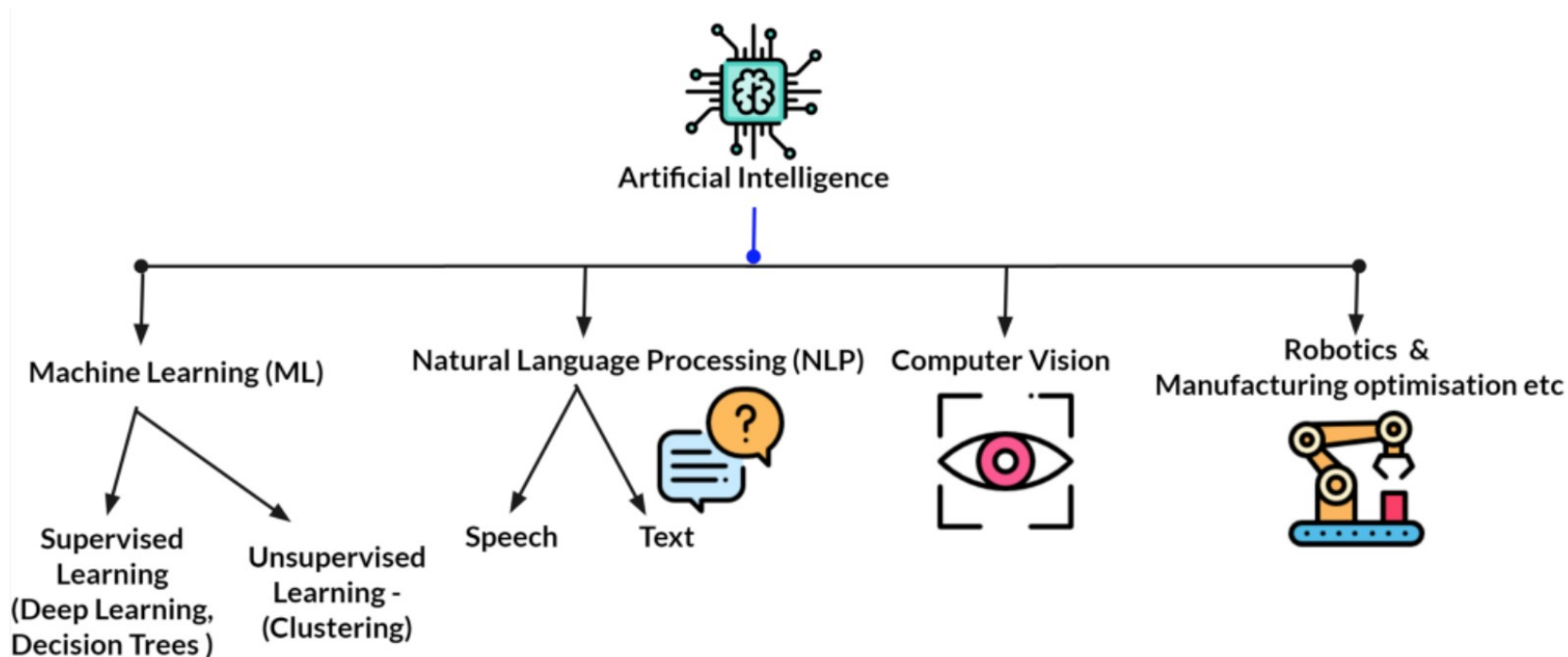


65 Bahasa





NLP and AI



Source : www.mc.ai

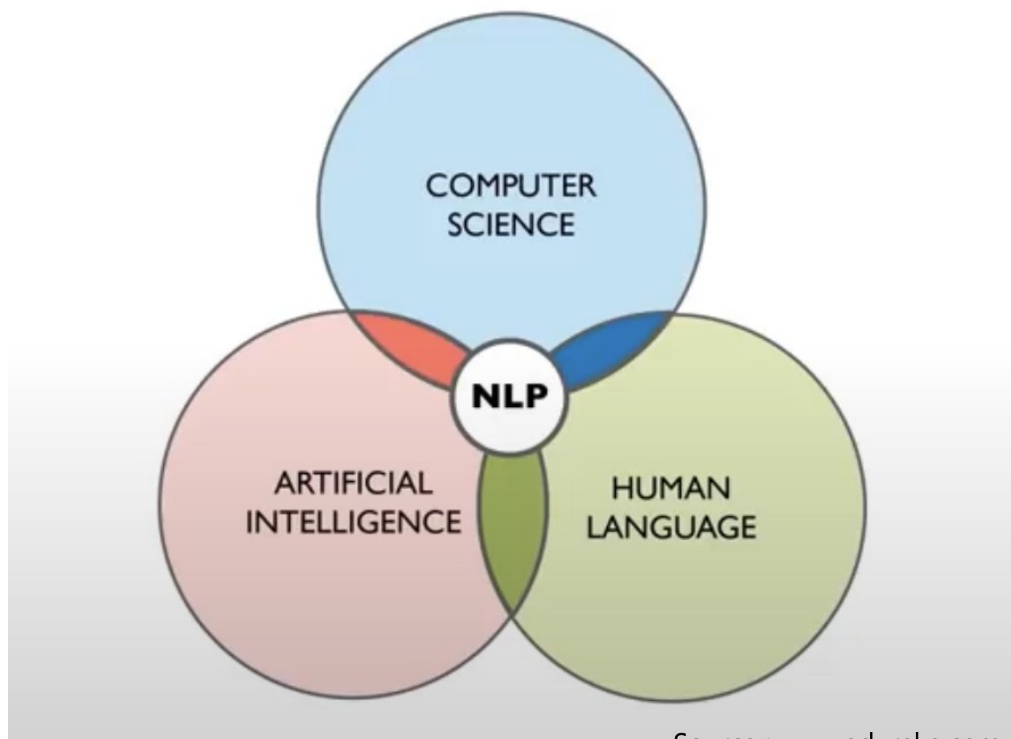


What is NLP ?

- A branch of AI that helps to understand, interpret and manipulate human language
- All about leveraging tools, techniques and algorithms to process and understand natural language based-data which is usually unstructured like text, speech, etc
- Sub-field of AI that is focused on enabling computers to understand and process human language
- NLP works based on how human use language (learn through experience)



What is NLP ?



Source : www.edureka.com



Brief History of NLP

- 1950s
 - Early MT : word translation +re-ordering
 - Chomsky's generative grammar
 - Bar-Hill's argument
- 1960-80s
 - Applications
 - BASEBALL : use NL interface to search database on baseball games
 - ELIZA : simulation of conversation with psychoanalyst
 - SHREDLU : use NL to manipulate block world
 - Message understanding : understand a newspaper article on terrorism
 - Machine translation
 - Methods
 - ATN (augmented transition networks) : extended context-free grammar
 - Case grammar (agent, object, etc)
 - DCG-Definite Clause Grammar
 - Dependency grammar : an element depends on another
- 1990s-now
 - Statistical methods
 - Speech recognition
 - MT system
 - Question-answering
 -



NLP examples

- Question Answering



kasus corona di indonesia



All Images News Maps Videos More Settings Tools

About 152,000,000 results (0.60 seconds)

Peringatan COVID-19

See results in English

Penyakit coronavirus (COVID-19)

Indonesia

Ringkasan

Gejala

Pencegahan

Perawatan

Statistik

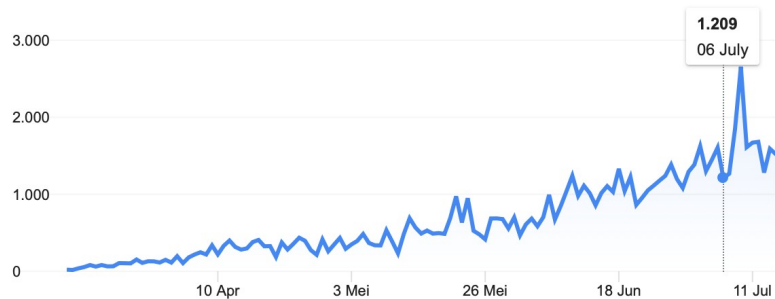
Bagikan

Perubahan harian

Kasus baru

Indonesia

Semua



Setiap hari menampilkan jumlah kasus baru yang dilaporkan sejak hari sebelumnya · Diperbarui kurang dari 2 hours ago · Sumber: [Wikipedia](#) · [Tentang data ini](#)

Peta penyebaran kasus



Sumber: [Wikipedia](#) dan [The New York Times](#) · [Tentang data ini](#)



NLP examples

- Machine Translation

☰ Google Translate

The screenshot shows the Google Translate interface. At the top, there are two tabs: 'Text' (selected) and 'Documents'. Below the tabs, the source language is 'INDONESIAN - DETECTED' and the target language is 'CHINESE (TRADITIONAL)'. The source text is 'Saya orang Indonesia' and the translated text is '我是印尼人'. Below the source text, there is a speaker icon and a character count '20/5000'. Below the translated text, there is a speaker icon and the pinyin 'Wǒ shì yìnní rén'.

NLP examples

- Sentiment Analysis



tirto.id

Q JELAJAH INDEPTH MILD REPORT CURRENT ISSUE



Donald John Trump

😊 43% 😐 5% 😞 52%



Prabowo Subianto Djojohadikusumo

😊 69% 😐 7% 😞 24%



Joko Widodo

😊 74% 😐 5% 😞 21%



Anies Rasyid Baswedan

😊 65% 😐 5% 😞 30%

PERSEPSI TERHADAP ISSUE

Lihat Semua

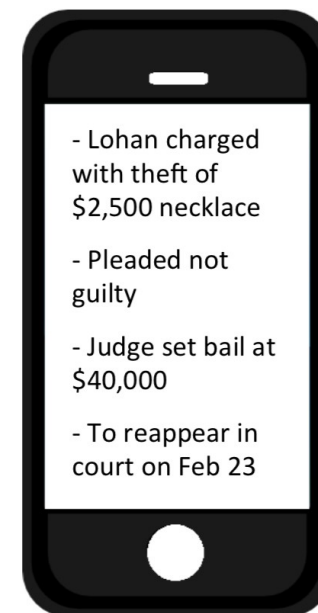
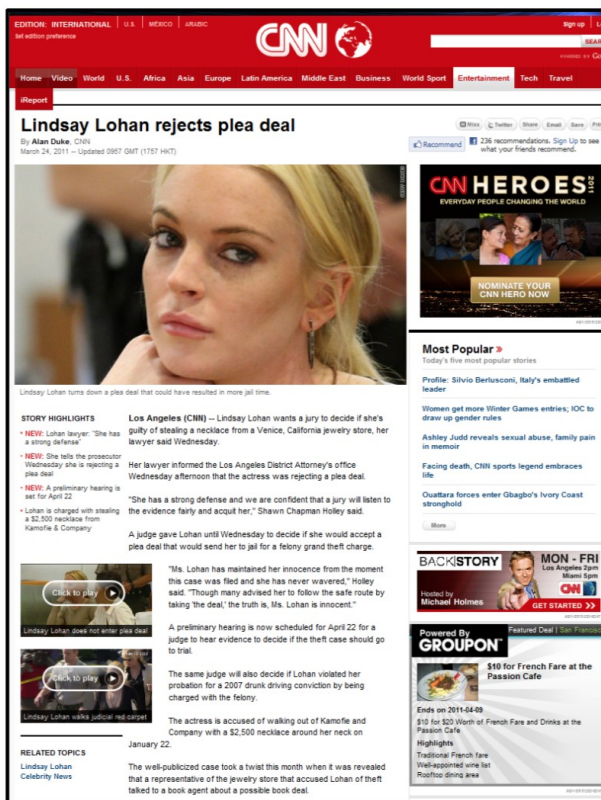
Issue	Persepsi Media	Persepsi Publik
Menkes Ganti Istilah ODP, PDP, OTG Corona	59% (Green), 8% (Grey), 33% (Red)	22% (Green), 49% (Grey), 29% (Red)
Bioskop Dibuka 29 Juli 2020	74% (Green), 7% (Grey), 19% (Red)	32% (Green), 57% (Grey), 11% (Red)
Solo Masuk Zona Hitam	54% (Green), 3% (Grey), 43% (Red)	10% (Green), 38% (Grey), 52% (Red)
Secapa TNI AD Jadi Klaster Corona	68% (Green), 5% (Grey), 27% (Red)	8% (Green), 75% (Grey), 17% (Red)
Penghentian Penerimaan CPNS 2 Tahun	79% (Green), 4% (Grey), 17% (Red)	5% (Green), 82% (Grey), 13% (Red)

powered by: newstensity.com



NLP Examples

- Summarization



<http://www.cs.unc.edu/~mbansal/>



NLP Examples

- Siri



Contains :

- Speech recognition
- Language analysis
- Dialog processing
- Text to speech



Component of NLP

Natural Language Understanding

- Mapping input to useful representation
- Analyzing different aspect of the language

Natural Language Generation

- Process of producing meaningful phrases and sentences in the form of natural language from some internal representation
- It involves :
 - Text Planning (retrieving the relevant content from knowledge base)
 - Sentence Planning (choosing required words, forming meaningful phrases, setting tone of the sentence)
 - Text Realization (mapping sentence plan into sentence structure)

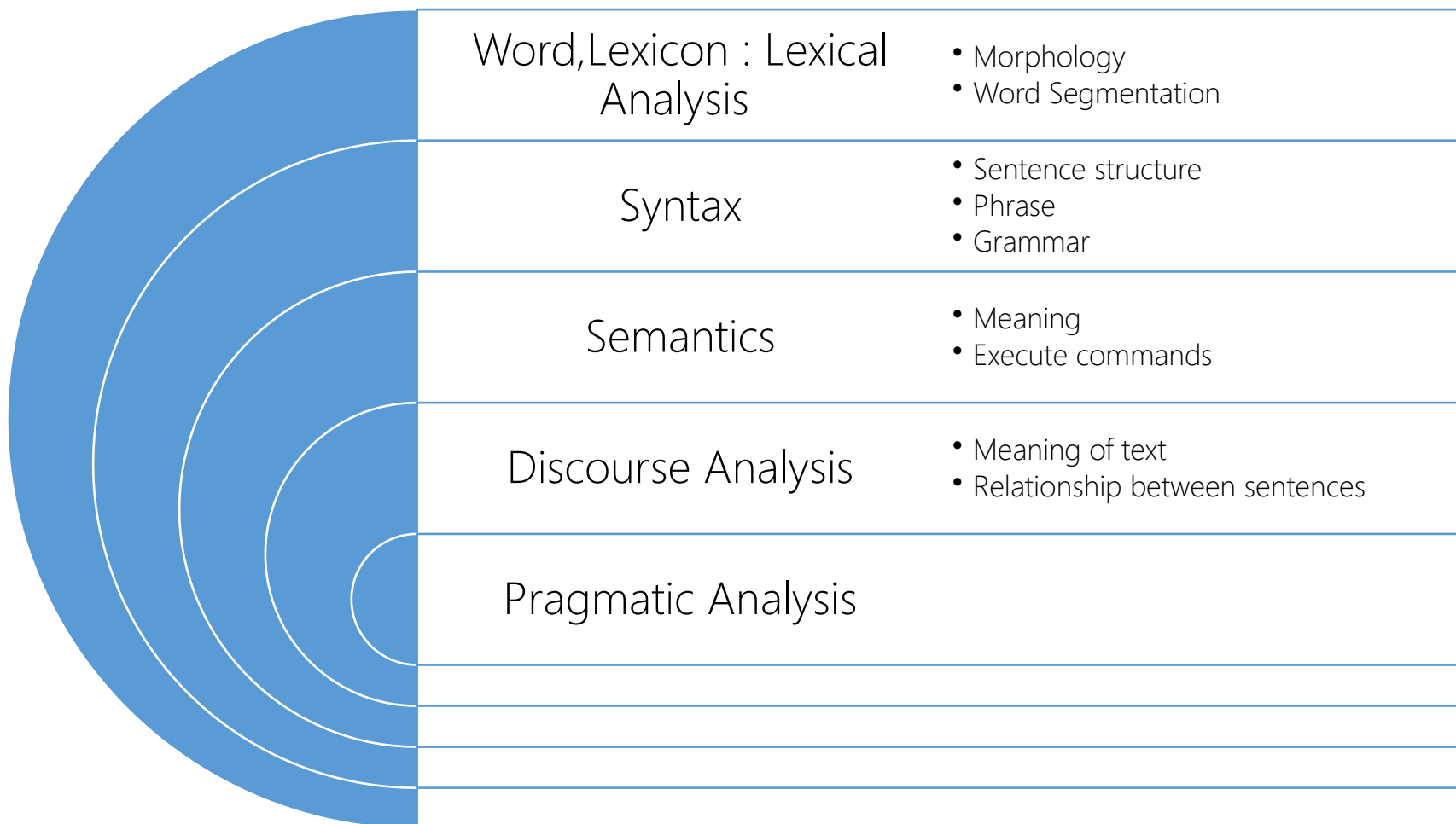


Problems in NLU

- Ambiguity
 - Lexical /Morphological : change (V,N), training(V,N), even(ADJ,ADV)...
 - Syntactic : Helicopter powered by human flies
 - Semantic : He saw a man on the hill with a telescope
 - Discourse : anaphora ,...
- Classical Solution
 - Using a later analysis to solve ambiguity of an earlier step
 - Eg. He gives him the change . (change as verb does not work for parsing)
 - He changes the places. (change as noun does not work for parsing)
 - However : He saw a man on the hill with a telescope
 - Correct multiple parsings
 - Correct semantics interpretations → semantic ambiguity
 - Use contextual information to disambiguate (does a sentence in the text mention that "He" holds a telescope?)



Aspects of Language Processing



NLP pipelines





Tokenization

- Break a complex sentence into words
- Understand the importance of each of the words with respect to the sentence
- Produce a structural description on an input sentence

Tokenization

is

the

first

step

in

NLP

Source : www.edureka.com



....Stemming

- Normalize words into its base form or root form

Affectation

Affects

Affections

Affected

Affection

Affecting

Source : www.edureka.com



....Lemmatization

Group together different inflected forms of a word, called Lemma

Somehow similar to Stemming, as it maps several words into one common root

Output of Lemmatization is a proper word

For example, a Lemmatizer should map gone, going and went into go

The difference with stemming is that a **stemmer** operates **without knowledge of context**, and therefore cannot understand the difference between words which have different meaning depending on parts of speech.
Lemmatization attempts to select the correct lemma depending on the context



....Lemmatization : Example

- The word “better” has “good” as its lemma. This link is missed by stemming, as it required a dictionary look-up
- The word “meeting” can be either the base form of a noun or a form of a verb (“to meet”) depending on the context ; e.g. “in our last meeting” or “we are meeting again tomorrow”



....Remove Stop Words

- Stop words are words which are **filtered out** before or after processing of text
- When applying machine learning to text, these words can add a lot of **noise**
- That's why we want to remove the irrelevant word
- Some example of stop words are **a, an, the, and** the like



POS : Parts of Speech

- POS are **specific lexical categories** to which words are assigned, based on their syntactic context and role

POS : Tags and Descriptions

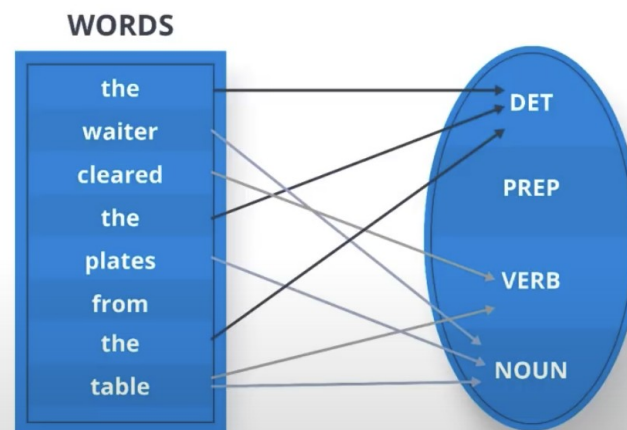
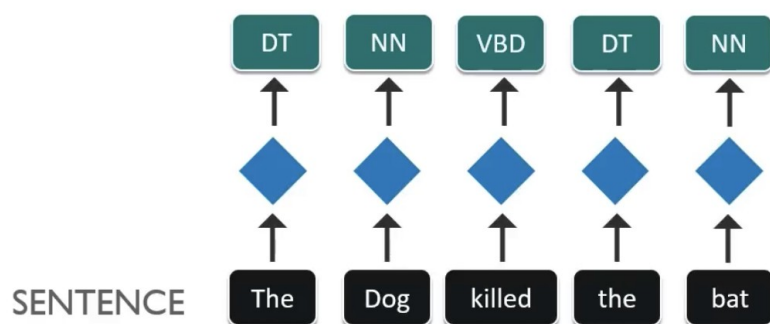
Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP\$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBR	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP\$	Possessive whpronoun
WRB	Whadverb

Source : www.edureka.com



POS : Examples

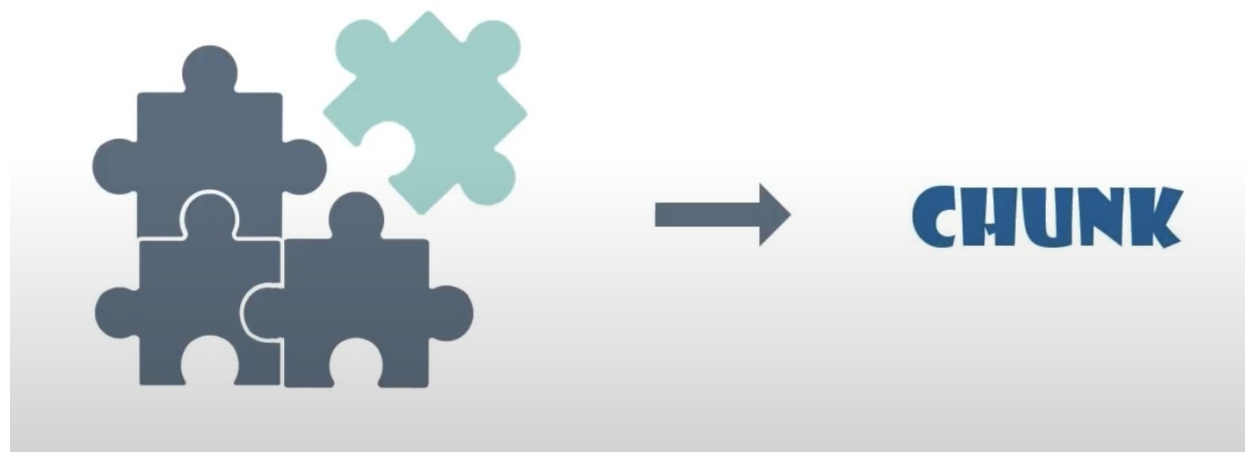


Source : www.edureka.com



Chunking

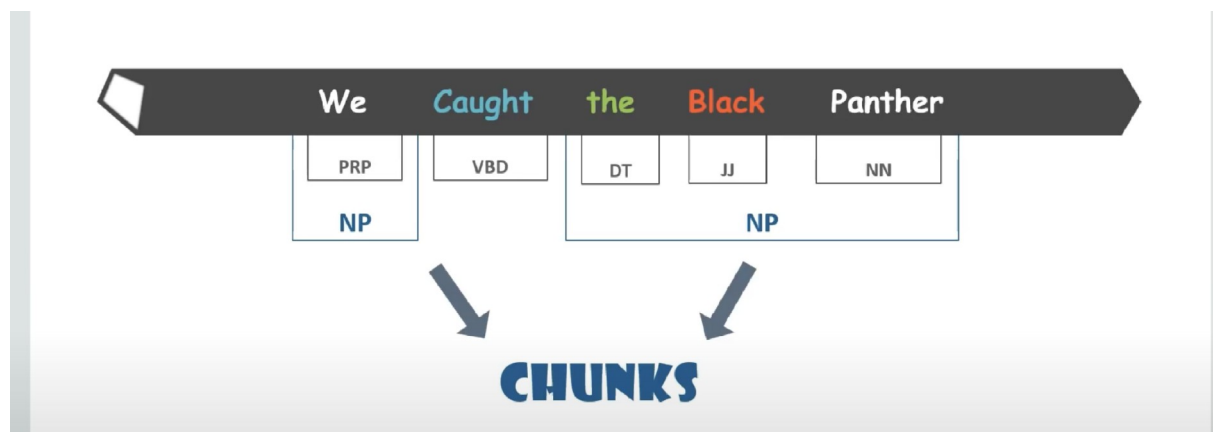
- Picking up Individual pieces of information and grouping them into bigger pieces



Source : www.edureka.com



Chunking : Example

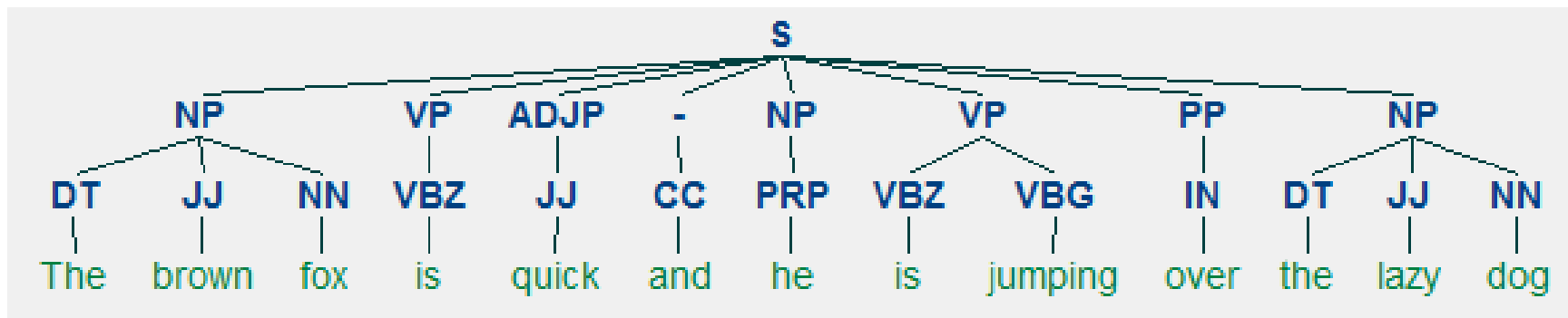


Source : www.edureka.com



Parsing

- Technique of analyzing the structure of a sentence to break it down into its smallest constituents (which are tokens such as words) and group them together into higher-level phrases.



Source : www.towardsdatascience.com

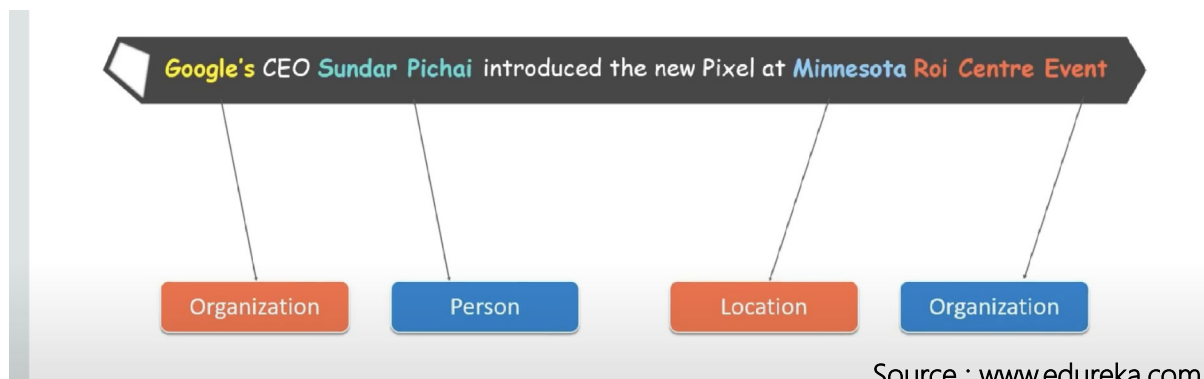


Named Entity Recognition (NER)

- A particular term that represent specific entities that are more informative and have unique context
- These entities are knows as named entities, which more specifically refer to terms that represent real-world object like people, place, organization, and so on, which are often denoted by proper name



NER : examples



US GPE unveils world's most powerful supercomputer, beats China GPE . The US GPE has unveiled the world's most powerful supercomputer called 'Summit', beating the previous record-holder China GPE 's Sunway TaihuLight ORG . With a peak performance of 200,000 CARDINAL trillion calculations per second ORDINAL , it is over twice as fast as Sunway TaihuLight ORG , which is capable of 93,000 CARDINAL trillion calculations per second. Summit has 4,608 CARDINAL servers, which reportedly take up the size of two CARDINAL tennis courts.

Source : www.towardsdatascience.com



How to implement NLP?

Machine Learning

- The learning NLP procedures used during machine learning
- It automatically focuses on the most common case

Statistical Inference

- NLP can make use of statistical inference algorithm
- It help us to produce model that are robust
 - e.g. containing words or structures which are known to everyone



Statistical Inference : example

- Topic Modeling
 - Type of statistical model for discovering the abstract “topics” that occur in collection of document
 - Frequently used text mining tool for discovery of hidden semantic structures in a text body
 - Topic models can help you automatically discover patterns in a corpus
 - **unsupervised** learning
 - Topic models automatically...
 - group topically-related words in “topics”
 - associate tokens and documents with those topics



Topic Model

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

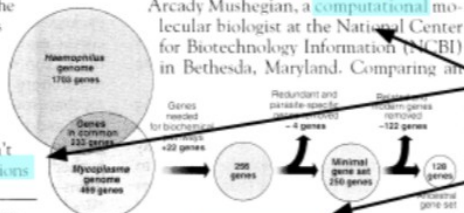
Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a postdoc at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

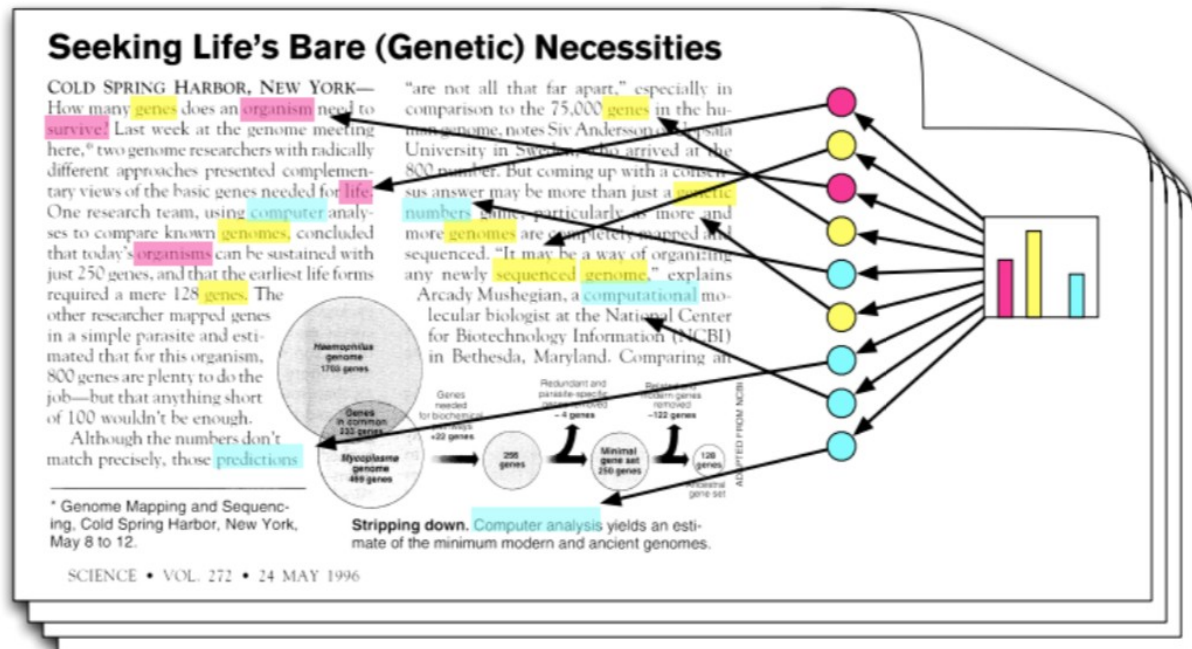


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



from David Blei



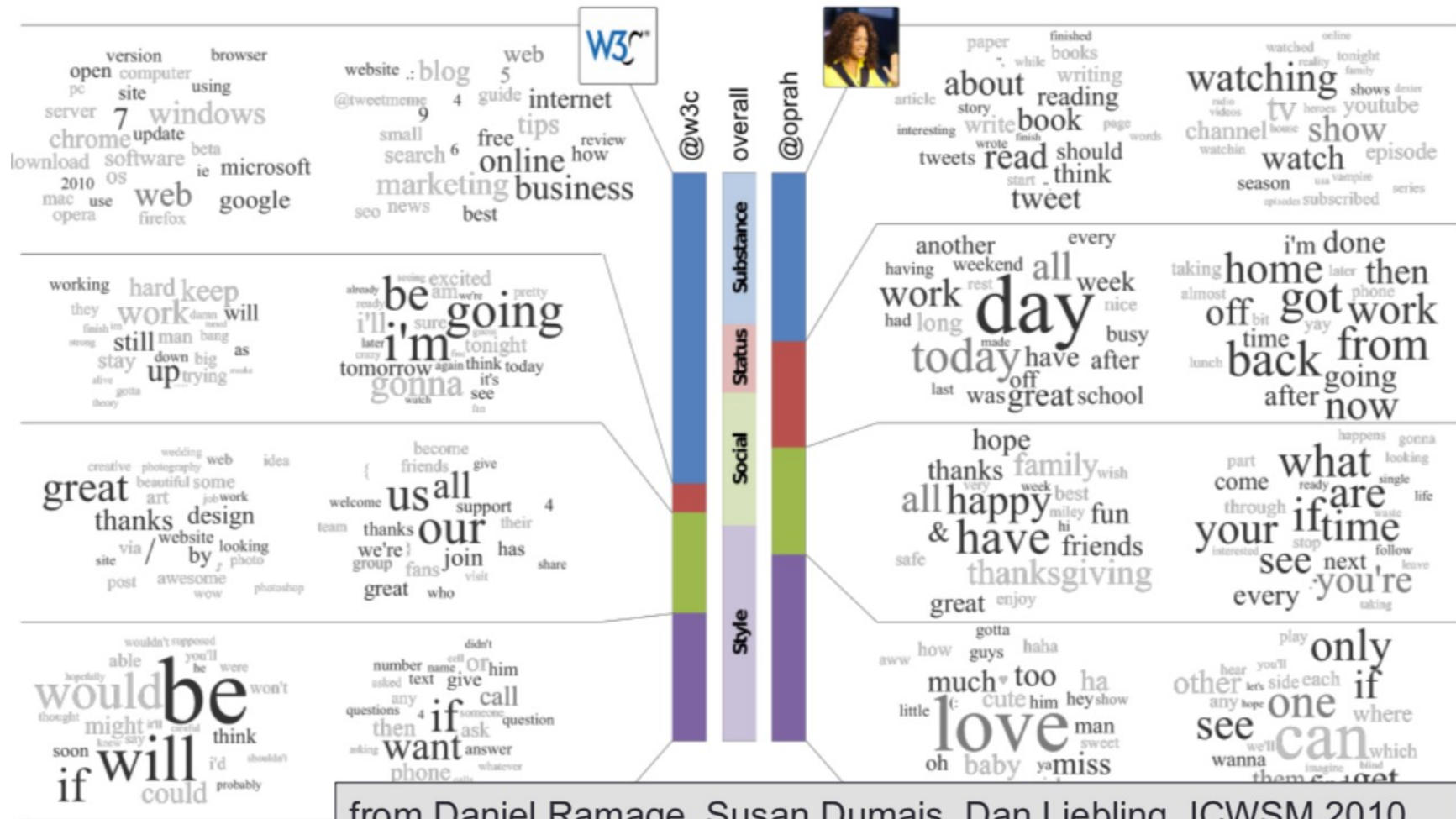
Implementation : Example

Suppose you want to learn something about a corpus that's too big to read

- What topics are trending today on Twitter?
 - What research topics receive grant funding (and from whom)?
 - What issues are considered by Congress (and which politicians are interested in which topic)?
 - Are certain topics discussed more in certain languages on Wikipedia?
- need to make sense of...
- **half a billion** tweets daily
 - **80,000** active NIH grants
 - **hundreds** of bills each year
 - **Wikipedia** (it's big)

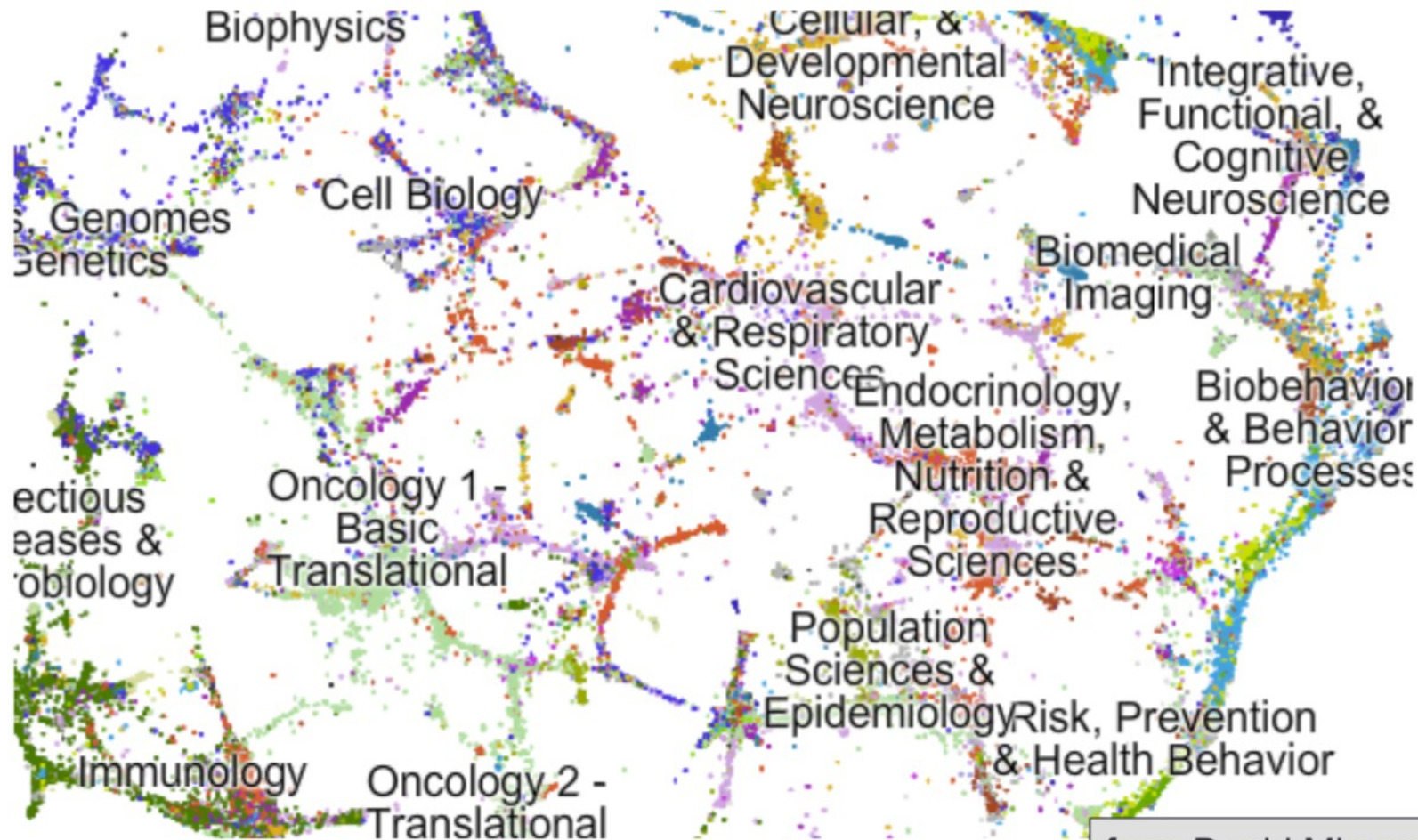


Twitter Topic





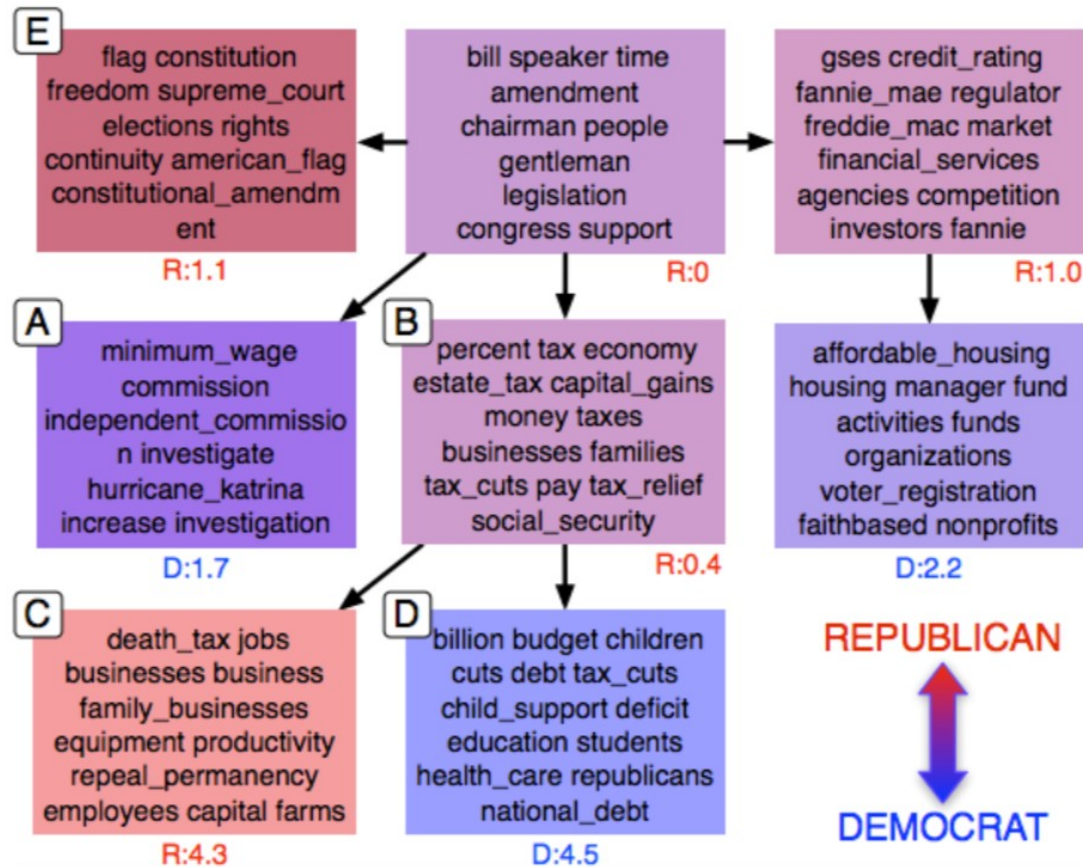
Research Grant Trends



from David Mimno



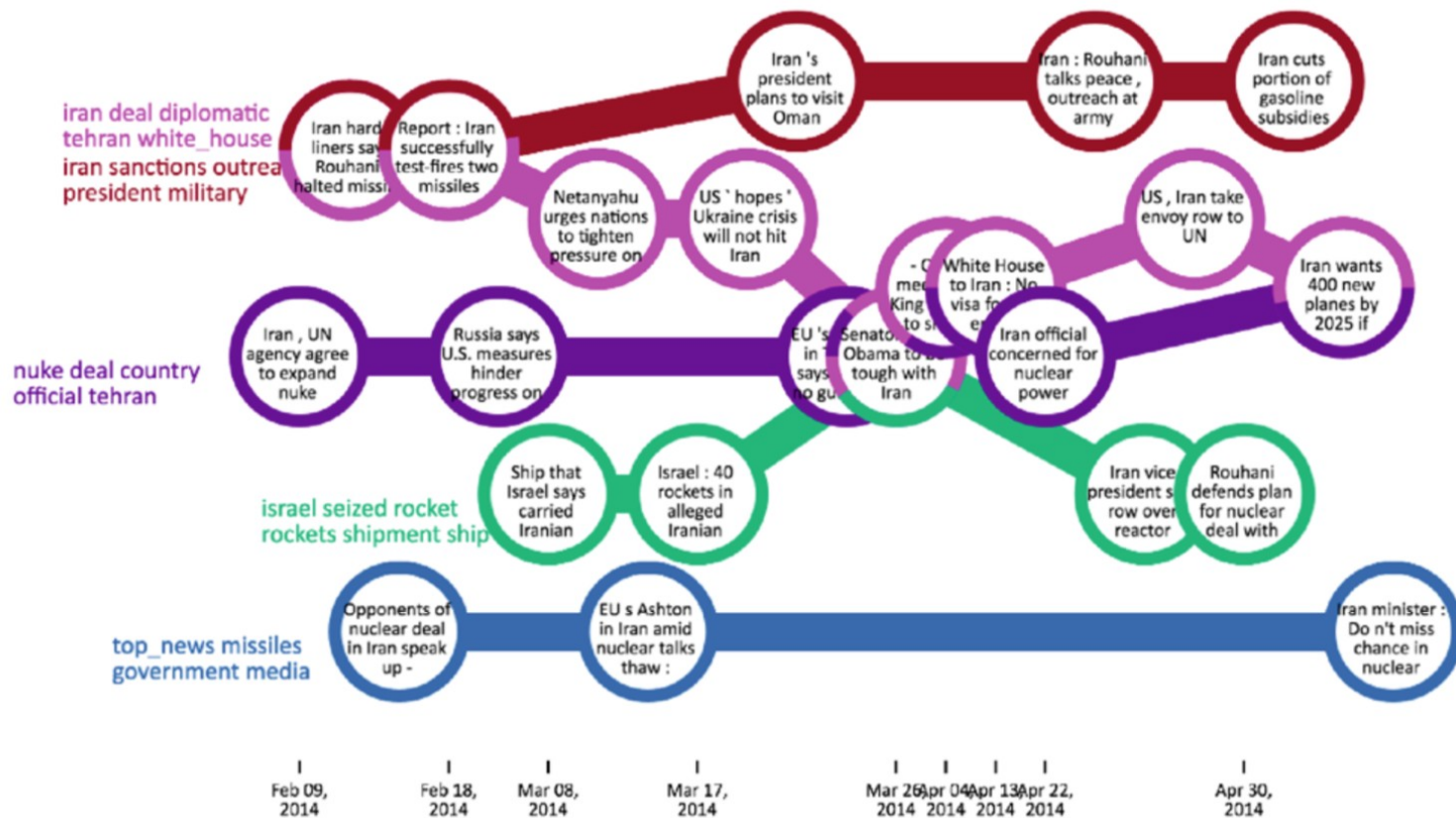
Political Issues



from Viet-An Nguyen, Jordan Boyd-Graber, Phillip Resnik. NIPS 2013.



Analyzing evolving stories in news



Roberto C.B, et,al, 2019



UNIVERSITAS
GADJAH MADA

Matur nuwun...
Thank you...
謝謝

