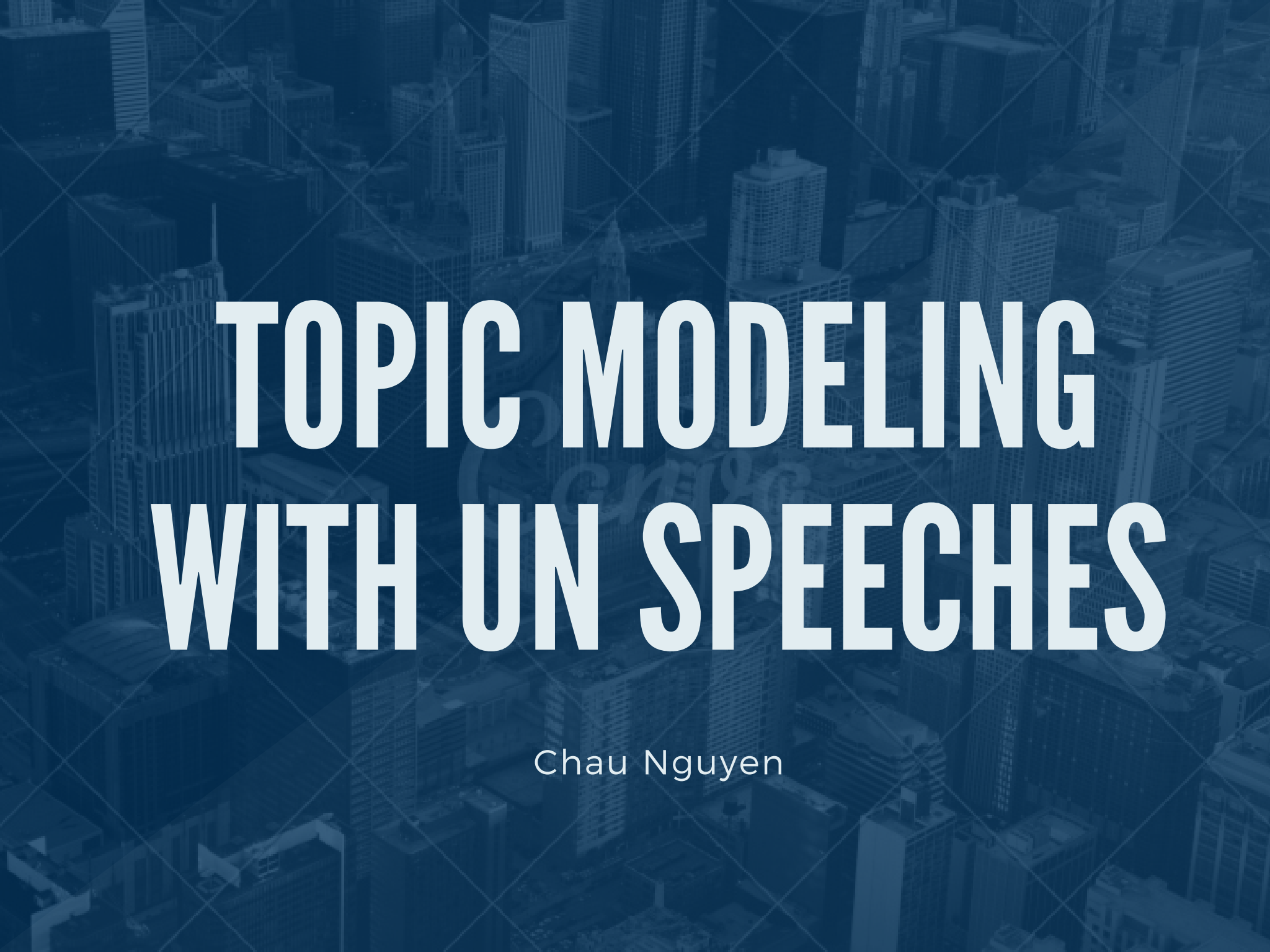


TOPIC MODELING WITH UN SPEECHES

Chau Nguyen

TOPIC MODELING - WHAT?

"In text mining, we often have collections of documents, such as blog posts or news articles, that we'd like to divide into natural groups so that we can understand them separately. Topic modeling is a method for **unsupervised classification** of such documents, similar to clustering on numeric data, which finds natural groups of items even when we're not sure what we're looking for. " - Julia Silge and David Robinson (authors of Text mining with R)

TOPIC MODELING - WHY?

- It takes manpower to go through documents to label them -> labor intensive and expensive
- Interpretable topics (sometimes)

TOPIC MODELING - HOW?

- Pre-processing text data
- Latent Dilrichlet Allocation
- Interpretation

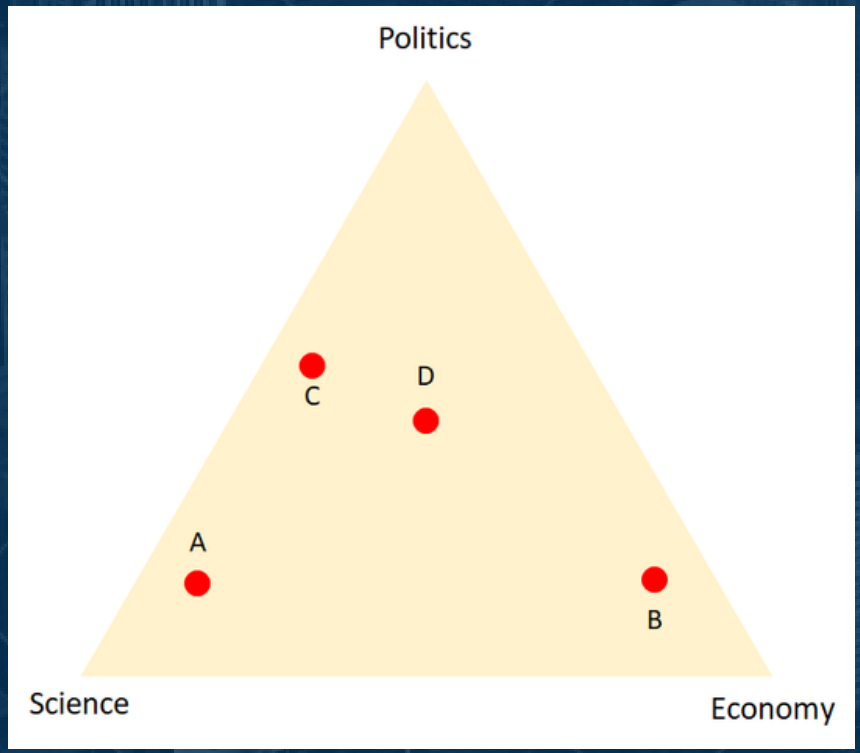
LATENT DIRICHLET ALLOCATION

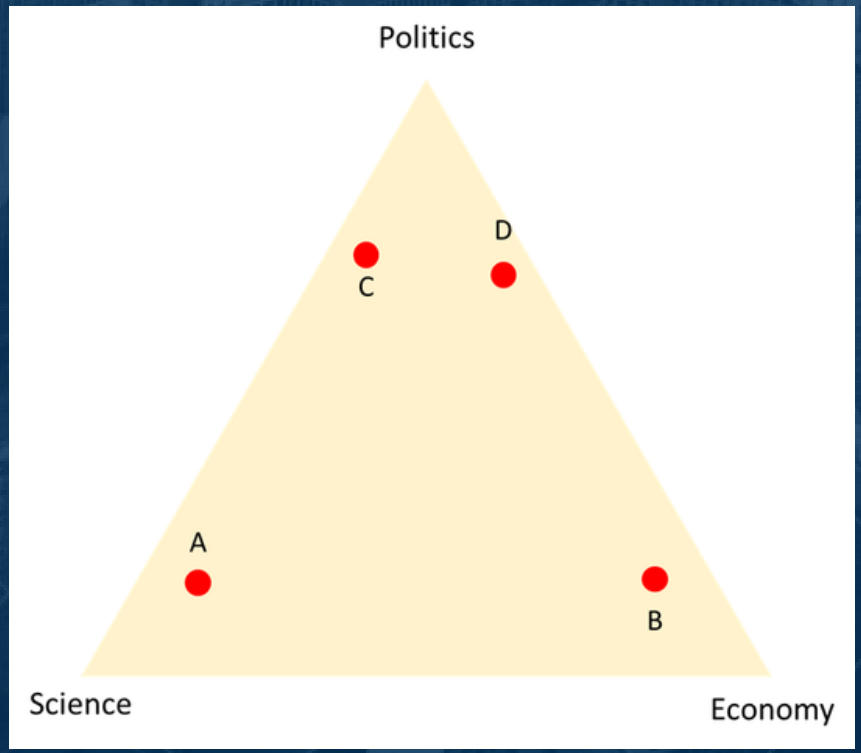
EVERY DOCUMENT IS A MIXTURE OF TOPICS.

EVERY TOPIC IS A MIXTURE OF WORDS.

DOCUMENT-TOPIC PROBABILITY

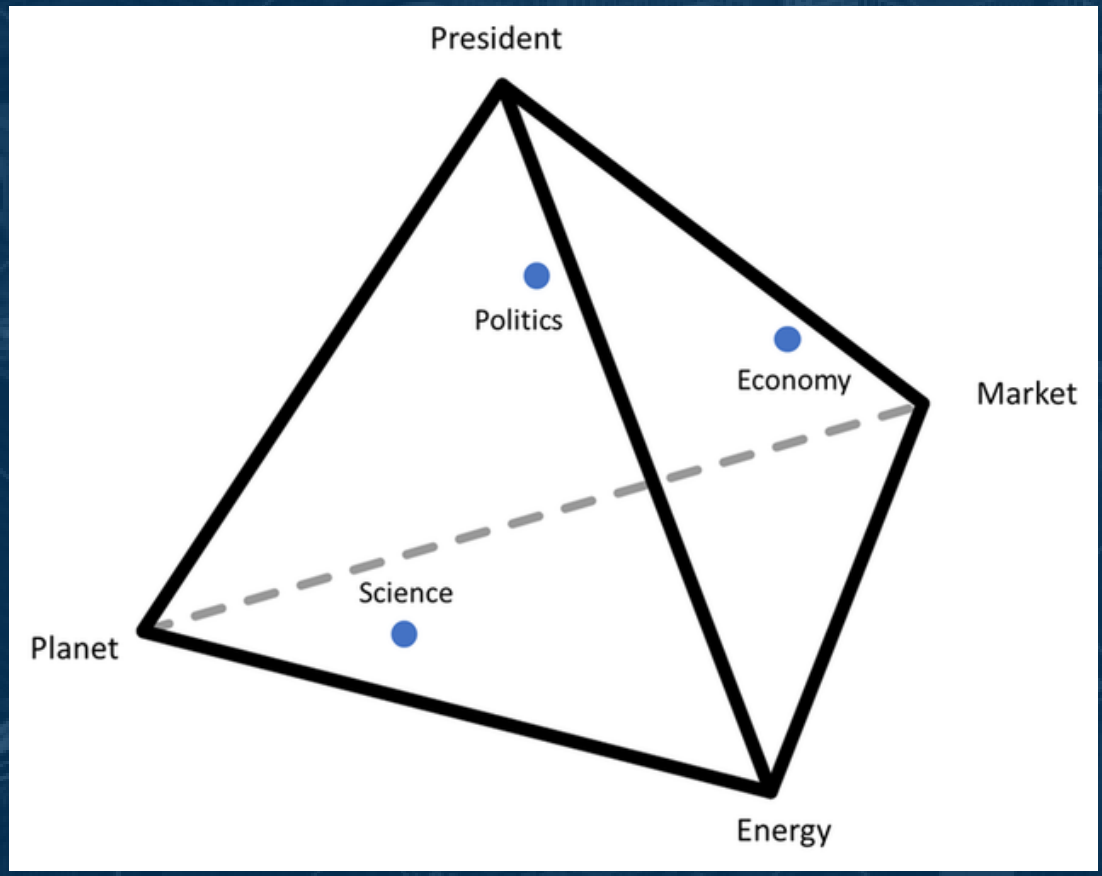
GOAL: MOVE EACH DOCUMENT CLOSER TO TOPICS AFTER EACH ITERATION





WORD-TOPIC PROBABILITY

IS THERE ANY TERMS THAT HAVE STRONG CONNECTION WITH CERTAIN TOPICS?



UN SPEECHES

- 45 YEARS, 199 COUNTRIES, 7507 DOCS

- PRE-PROCESSING
  - Remove line breaks, tabs, double white space
  - Replace - and _ with white space
  - Remove sentence index number
  - Remove common stop words (a, an, the, etc)

APPLY LDA ON UN SPEECHES

CHOOSE 5 TOPICS:

TOPIC 1: 'development', 'nations', 'united', 'international', 'countries'
TOPIC 2: 'nations', 'united', 'international', 'world', 'security'
TOPIC 3: 'united', 'people', 'states', 'countries', 'world'
TOPIC 4: 'united', 'people', 'nations', 'peace', 'world'
TOPIC 5: 'international', 'countries', 'world', 'nations', 'peace'

INTERPRETATION?

1: something about the world
2: something about the world
3: something about the world
4: something about the world
5: something about the world

NOT VERY HELPFUL!!!

WHY?

TOP 50 WORDS IN UNITED NATIONS SPEECHES

NEED TO FILTER OUT COMMON WORDS

MAX_DF PARAMETER:

- Remove terms that appear in X % of documents.
- We need collections of terms that is common enough and shared by several documents to indicate a shared topics/latent information but also unique enough that it is not shared by all documents.

NEED TO FILTER OUT COMMON WORDS

LDA WITH MAX_DF = 0.50

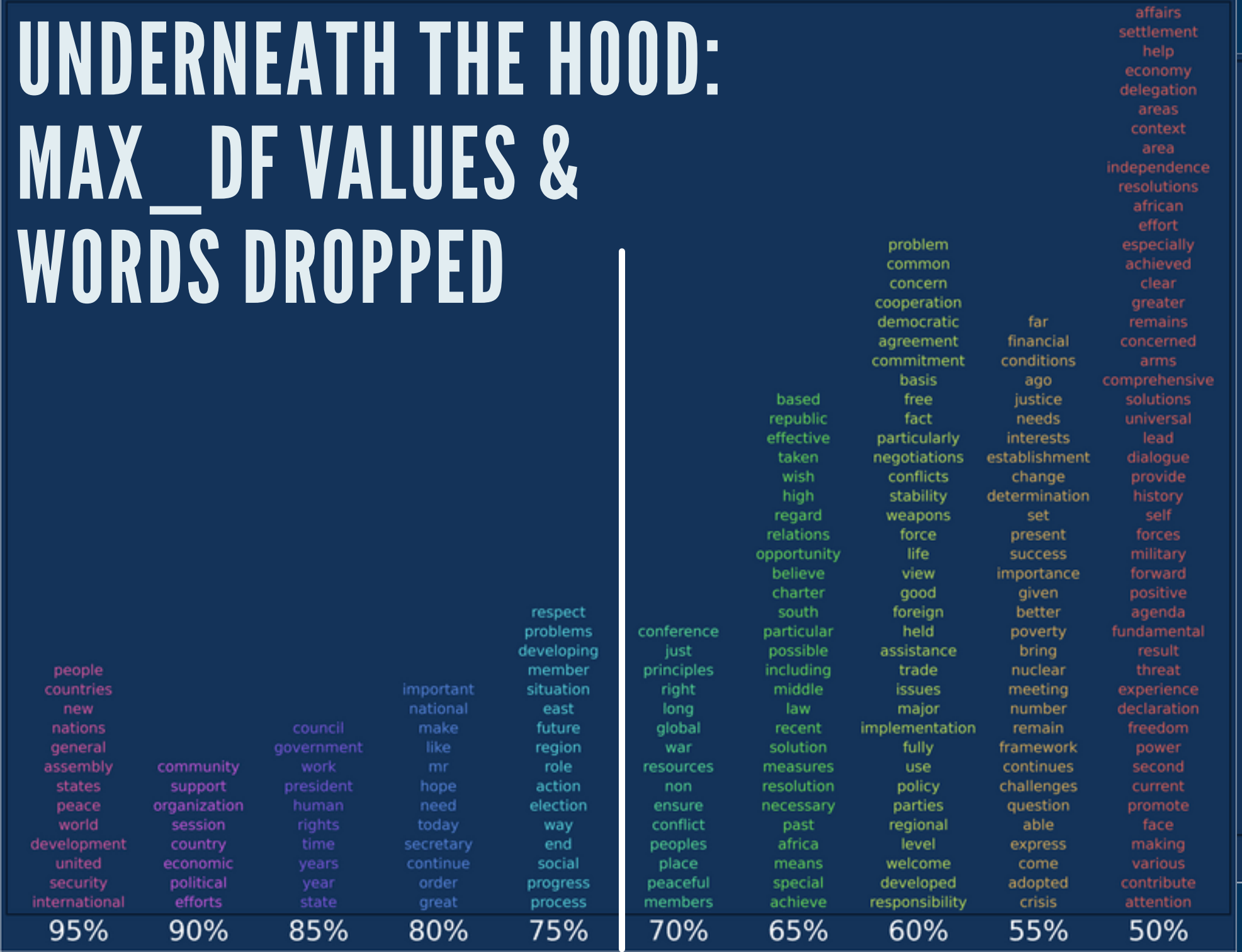
PICK 5 TOPICS:

Central American debt?: 'debt', 'environment', 'democracy', 'continent', 'central'
Latin/ Central America: 'america', 'american', 'latin', 'central', 'democracy'
TOPIC 3: 'operation', 'disarmament', 'namibia', 'powers', 'struggle'
Middle east: 'arab', 'israel', 'palestinian', 'iraq', 'israeli'
Climate change: 'sustainable', 'climate', 'reform', 'small', **'terrorism'**

REDUCE TO 4 TOPICS

Climate change: 'reform', 'sustainable', 'climate', 'environment', 'goals'
Latin/ Central America: 'america', 'american', 'small', 'latin', 'democracy'
TOPIC 3: 'operation', 'disarmament', 'powers', 'struggle', 'namibia'
Middle East: 'arab', 'israel', 'iraq', 'palestinian', 'terrorism'

UNDERNEATH THE HOOD: MAX_DF VALUES & WORDS DROPPED



SET MAX_DF CUTOFF AT 0.75

59 WORDS DROPPED

				respect
				problems
				developing
people				member
countries			important	situation
new			national	east
nations		council	make	future
general		government	like	region
assembly	community	work	mr	role
states	support	president	hope	action
peace	organization	human	need	election
world	session	rights	today	way
development	country	time	secretary	end
united	economic	years	continue	social
security	political	year	order	progress
international	efforts	state	great	process
95%	90%	85%	80%	75%

FURTHER CONSIDERATIONS

MODIFY STOP_WORDS

Added the 59 words to sklearn's 'english' stop words

SUBSET FOR COUNTRIES/ YEARS

Created function to quickly perform LDA based on country/ year subset

WHEN SUBSETTING COUNTRY, NEED ADDITIONAL LAYER OF MAX_DF

Why?

FURTHER CONSIDERATIONS

VIETNAM, MAX_DF = 1.0

[['south', 'asia', 'kampuchea', 'nam', 'viet'],
 ['nam', 'viet', 'cooperation', 'relations', 'asean'],
 ['viet', 'nam', 'peoples', 'soviet', 'war'],
 ['viet', 'nam', 'cambodia', 'cambodian', 'cooperation'],
 ['viet', 'nam', 'peoples', 'kampuchea', 'independence']]

VIETNAM, MAX_DF = 0.95

[['kampuchea', 'struggle', 'war', 'vietnamese', 'republic'],
 ['issues', 'nation', 'reform', 'view', 'century'],
 ['terrorism', 'law', 'environment', 'reform', 'view'],
 ['poverty', 'century', 'china', 'developed', 'declaration'],
 ['cambodia', 'view', 'nation', 'mutual', 'concerned']]

TO SUBSET OR NO?

45 YEARS OF SPEECHES, 199 COUNTRIES

- Too much data for meaningful topic modeling
- Each runthrough is computationally expensive

SUBSET DEPENDING ON USE-CASE

- What are all countries in the world talking about?
- What is each individual country talking about?

LDA PARAMETERS DEPEND ON SUBSET

- More documents -> less n_component

WORLD 15-YEAR INTERVALS

5 TOPICS

1970-1985

International law: sea, law, nuclear, disarmament, power
WMD: nuclear, disarmament, global, weapons, settlement
Africa: african, certain, struggle, solidarity, justice
Middle East: israel, arab, america, palestinian, american
Cold War: struggle, nuclear, soviet, aggression, forces

1986-2000

WMD: cooperation, nuclear, global, weapons, reform
Middle East: africa, african, arab, israel, palestinian
African debt: africa, operation, debt, problem, trade
Democracy: democracy, small, century, poverty, democratic
Cold War: nuclear, republic, military, afghanistan, soviet

2001-2015

Africa: africa, african, republic, conflict, poverty
Terrorism: terrorism, european, nuclear, law, afghanistan
Climate changeclimate, change, small, sustainable, island
Middle East: palestinian, arab, israel, iraq, terrorism
Poverty: poverty, peoples, cent, democracy, america

SINGAPORE

5-YEAR INTERVALS, 5 TOPICS

2000-2005

9/11: 'september', 'common', 'key', 'attacks', 'consensus'
China/Taiwan: 'china', 'rules', 'trade', 'chinese', 'taiwan',
Islam: 'muslims', 'muslim', 'non', 'islamic', 'religious'
IIraq war: 'iraq', 'interests', 'power', 'members', 'charter'
Millenium development goals: 'developed', 'peacekeeping', 'millennium', 'trade', 'summit'

2006-2010

SEA: 'regional', 'myanmar', 'asean', 'common', 'powers'
SEA: 'asean', 'regional', 'cooperation', 'european', 'organizations'
Development: 'knowledge', 'education', 'urban', 'restructuring', 'cities'
GFC: 'economy', 'financial', 'crisis', 'governments', 'meeting'
Islam:'muslims', 'muslim', 'non', 'islamic', 'religious'

2011-2015

Development: 'urban', 'knowledge', 'education', 'post', 'management',
??: 'firmly', 'forum', 'effective', 'provide', 'best'
??: 'firmly', 'forum', 'effective', 'provide', 'best'
Terrorism: 'post', 'isis', 'cooperation', 'conference', 'south'
International cooperation: 'food', 'trade', 'issues', 'groupings', 'multilateral'

APPLICATIONS TO PUBLIC POLICY

SEMI-SUPERVISED

- Still need humans to check and make sense of the topics

LABEL ARCHIVED TEXTS

KEEP TRACK OF PUBLIC OPINIONS