

Utiliser les **connaissances** du  
**sens commun** pour la découverte de  
sujets **interprétables**

Ismail Harrando, Raphaël Troncy



**CIMPLE**

# What is Topic Modeling?

## Osaka says US Open win has given her belief. 2019-01-03

Japanese star Naomi Osaka said her US Open victory in September had given her the self-belief to be able to come from behind and win tight matches. Osaka was speaking after recovering from losing the first set of her Brisbane International quarter-final to Latvia's Anastasija Sevastova on Thursday. The world number five overhauled Sevastova to win 3-6, 6-0, 6-4 and reach the semi-finals of the season-opening tournament. "I feel like right now I'm really confident in myself, and I feel like the off-season training that I've been doing is really paying off," she said. "And I'm not sure if I would have had the same feeling six months ago. Six months ago I was feeling like I was not ready for the US Open." Osaka's stunning victory over the world number one in New York's Flushing Meadows in

SPORT

## World's oceans are heating up at a quickening pace: study. 2019-01-10

The world's oceans are heating up at an accelerating pace as global warming threatens a diverse range of marine life and a major food supply for the planet, researchers said Thursday. The findings in the US journal Science, led by the Chinese Academy of Sciences, debunk previous reports that suggested a so-called hiatus in global warming in recent years. The study says such hiatus ever existed, and its effect on the planet's main buffer -- the oceans. "Ocean heating is a very important indicator of climate change, and we have robust evidence that it is warming more rapidly than we thought," said co-author Zeke Hausfather, a graduate student in the Energy and Resources Group at the University of California, Berkeley.

ENVIRONMENT

# Topic Modeling

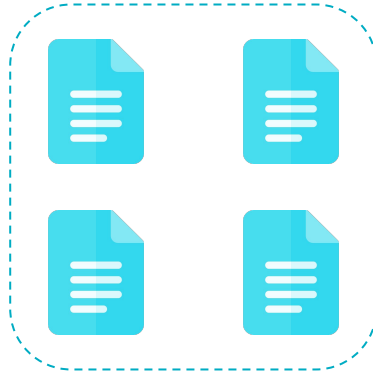


# Topic Modeling



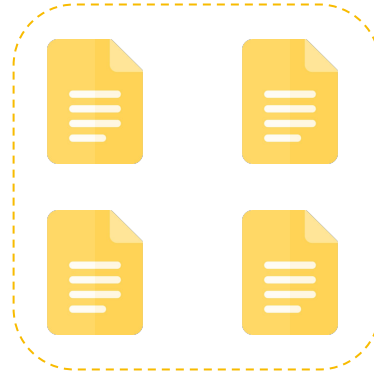
## “Sport”

game  
score  
player  
match  
team



## “Technology”

computing  
hardware  
computer  
digital  
internet



## “Politics”

politician  
parliament  
president  
government  
election

# Using Topic Models

## NLP Tasks

- Document retrieval
- Keyword Extraction
- Text Classification
- ..

## Data Exploration

- Visualization
- Interpretation
- Corpus Analysis
- ..

Accuracy  
Precision  
Ranking  
..

Coherence?

# Different methods

## STATISTICAL

Latent Dirichlet  
Allocation

**LDA** *Blei et al., 2003*

Hierarchical Dirichlet  
Process

**HDP** *Teh et al., 2006*

Gibbs Sampling  
for a DMM

**GSDMM** *Yin and Wang, 2014*

Latent Semantic  
Indexing

**LSI**

*Deerwester et al., 1990*

Non-negative Matrix  
Factorisation

**NMF**

*Paatero and Tapper, 1994*

## LINEAR ALGEBRA

Latent Feature  
Topic Models

**LFTM**

*Nguyen et al., 2015*

Contextualized  
Topic Model

**CTM**

*Bianchi et al., 2020*

## EMBEDDING-BASED

Paragraph Vector  
Topic Model

**PVTM**

*Lenz and Winker, 2020*

Distributed  
Representations  
of Topics

**Topic2Vec**

*Niu and Dai, 2015*

**Top2Vec**

*D. Angelov, 2020*

**BERTopic**

*M. Grootendorst, 2020*

doc2topic

**D2T**

[https://github.com/sronnqvist/](https://github.com/sronnqvist/doc2topic)

[doc2topic](https://github.com/sronnqvist/doc2topic)

## NEURAL

# Evaluating Topic Models

## **Reading Tea Leaves: How Humans Interpret Topic Models**

*NeurIPS, 2009*

## **Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence**

*NeurIPS, 2021*

## **Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures**


*NAACL, 2021*

## **Apples to Apples: A Systematic Evaluation of Topic Models**

*RANLP, 2021*

  
EURECOM

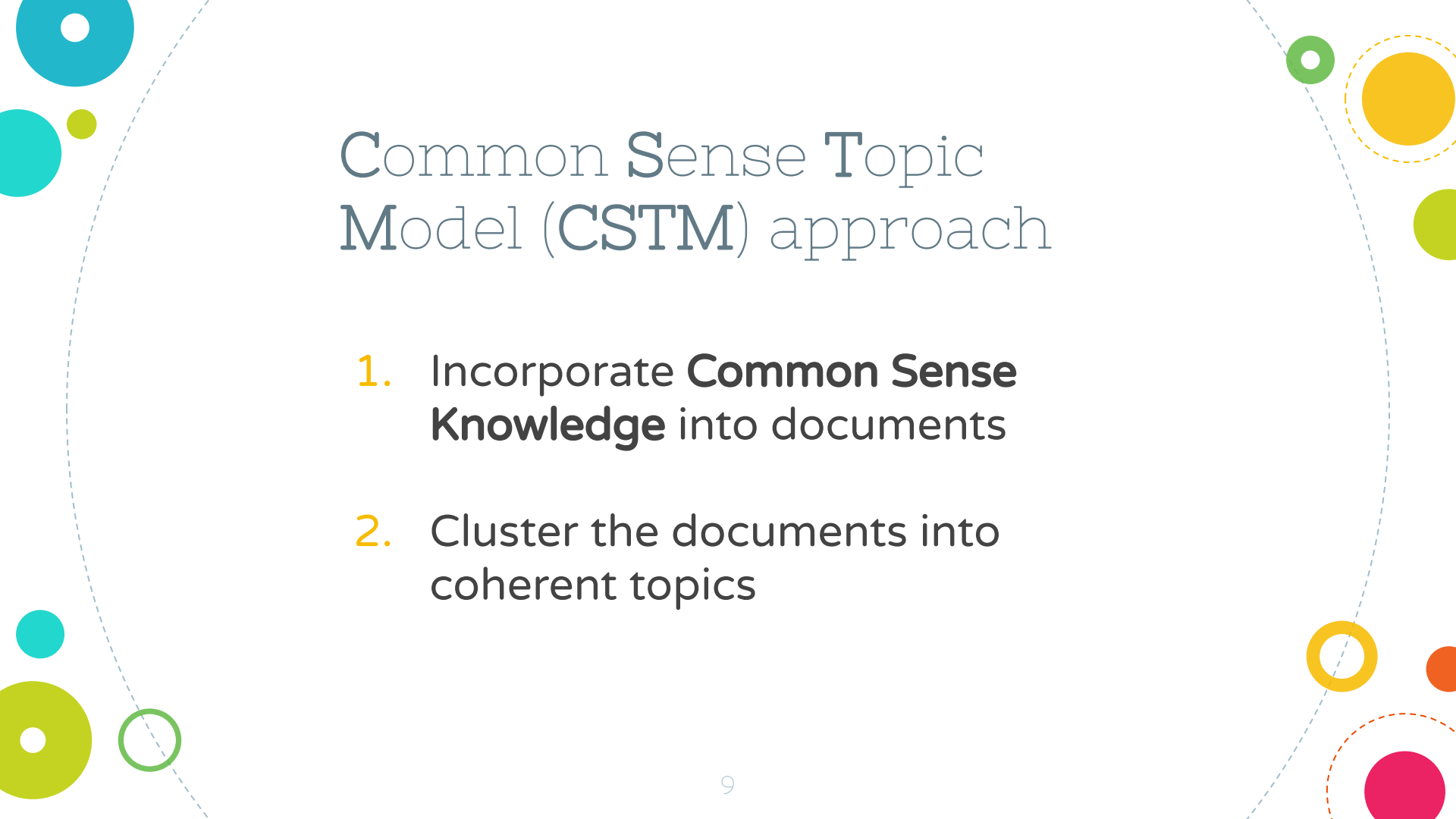
**Coherence?**

A decorative graphic consisting of a large, light blue dashed circle that frames the central text. Various colored circles (solid and hollow) in shades of teal, green, yellow, orange, and pink are scattered around the perimeter of the dashed circle.

How to make **topics** more interpretable?

Common sense as  
Domain Knowledge!



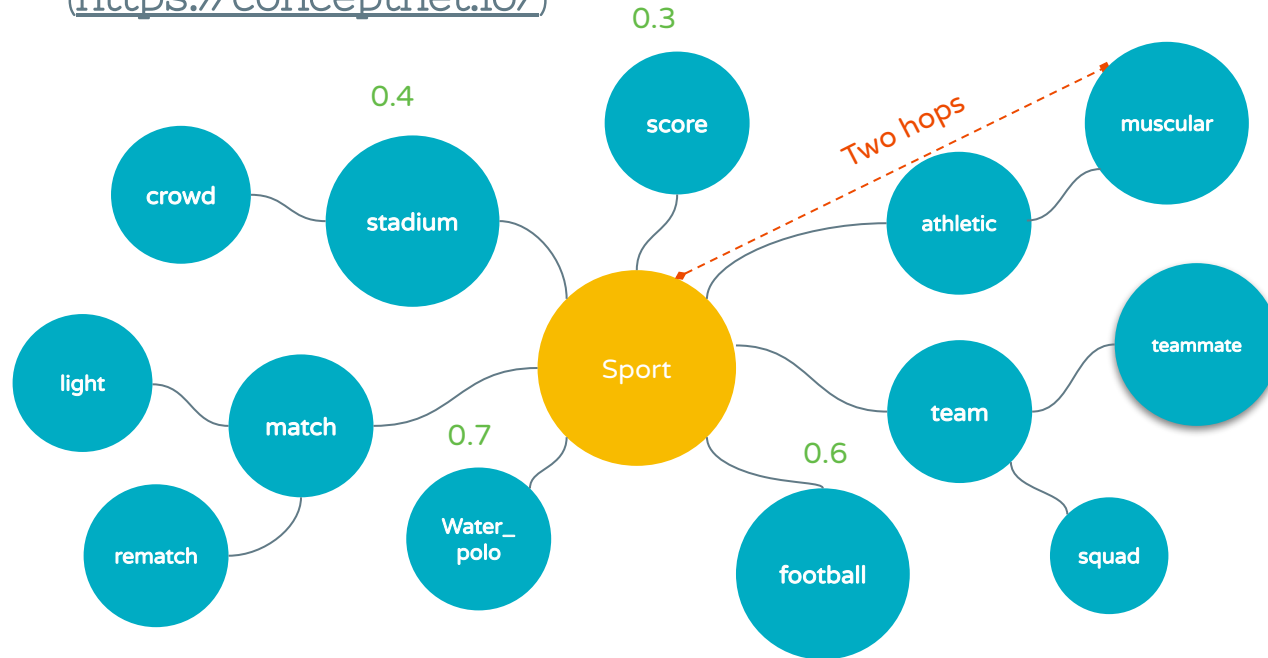
A decorative graphic consisting of a large, light blue dashed circle that frames the central text. Various colored circles (teal, yellow, green, orange, pink) are scattered around the perimeter of the dashed circle, some overlapping it. The colors include teal, yellow, green, orange, and pink.

# Common Sense Topic Model (CSTM) approach

1. Incorporate **Common Sense Knowledge** into documents
2. Cluster the documents into coherent topics

# Common-Sense Knowledge from ConceptNet

(<https://conceptnet.io/>)



**Conceptnet 5.5: An open multilingual graph of general knowledge**

R. Speer, J. Chin, and C. Havasi. - *Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017

# 1. Common-Sense enhanced Bag of Words (CS-BoW)

## Document

“ French team wins **football** championship ”

ball	0
championship	1
clown	0
friend	0
football	1
french	1
game	0
match	0
pain	0
peace	0
sport	0
team	1
telephone	0
war	0
win	1

BoW

+\*



*\* for each word in the document, we generate a filtered subgraph*

ConceptNet

=

ball	1
championship	1
clown	0
friend	0
football	1
french	1
game	0
match	1
pain	0
peace	0
sport	1
team	1+1
telephone	0
war	0
win	1+1

CS-BoW

## 2. CSTM



Corpus



CS-BoW

K-Means



N = 3



Clusters / Topics



# Evaluation

# Datasets

## → **BBC News:**

A news dataset from BBC containing 2225 English news articles classified in 5 categories: "Politics", "Business", "Entertainment", "Sports" and "Tech"

## → **AG News:**

A news dataset containing 127600 news articles from various sources, fairly distributed among 4 categories: "World", "Sports", "Business" and "Sci/Tech"

## → **20NewsGroup:**

a collection of 18000 user-generated forum posts arranged into 20 groups seen as topics such as "Baseball", "Space", "Cryptography", and "Middle East".

## → **AFP News:**

a dataset containing 70K English news articles issued by the French News Agency, with top-level categories such as "Politics", "Art, Culture and Entertainment", "Environment". The label distribution is highly unbalanced.

# Quantitative Analysis

Dataset	Model	V-Measure	WE-Coherence	NPMI
BBC	CSTM	<b>0.789</b>	<b>0.382</b>	-0.132
	K-Means	0.662	0.346	0.105
	LDA	0.729	0.359	0.122
	NMF	0.172	0.371	0.0225
AG News	CSTM	0.250	<b>0.387</b>	-0.0539
	K-Means	0.171	0.225	<b>0.027</b>
	LDA	<b>0.542</b>	0.214	0.001
	NMF	0.092	0.306	-0.0017
20NG	CSTM	0.403	0.303	-0.055
	K-Means	<b>0.433</b>	0.246	<b>0.127</b>
	LDA	0.403	<b>0.353</b>	0.031
	NMF	0.274	0.281	0.092
AFP	CSTM	0.431	0.296	-0.0459
	K-Means	<b>0.447</b>	<b>0.329</b>	<b>0.159</b>
	LDA	0.297	0.322	0.075
	NMF	0.409	0.308	0.127

Quantitative performance of CSTM and Baselines on 4 datasets.  
Best result on each dataset-metric pair is highlighted in bold

# Human evaluation

## → Word Intrusion

e.g. game, stock, sport, football, rugby

> Intruder word is :

**stock**

## → Topic Labeling

e.g. medicine, illness, disease, medication, medical

> Corresponding label is :

**Medicine**

## → Topic Classification

e.g. “Deutsche Bank reports first annual profit in four years”

> Corresponding topic is :

**business, commerce, value, market, finance**



# Quantitative Analysis

Model	Word Intrusion	Topic Labeling	Topic Classification
<b>CSTM</b>	<b>83.3%</b>	<b>84.6%</b>	<b>27.5%</b>
<b>K-Means</b>	33.3%	81.7%	19.5%
<b>LDA</b>	29.2%	52.9%	13.3%

Scores percentage (w.r.t the maximum obtainable)  
across datasets for CSTM, K-Means and LDA

# Future Work

- Try other BoW/TF-IDF variants for CS-BoW
- Use other clustering methods (even topic models!)
- Experiment with other CS Knowledge Graphs
- Impact of number of topics/hyperparameters
- Towards topic labeling

Thank you  
Any questions?

K-CAP 2021: <https://doi.org/10.1145/3460210.3493586>



**E-mail:** [ismail.harrando@eurecom.fr](mailto:ismail.harrando@eurecom.fr)

**Twitter:** @harrando

**Code & Evaluation:**

<https://github.com/D2KLab/CSTM>