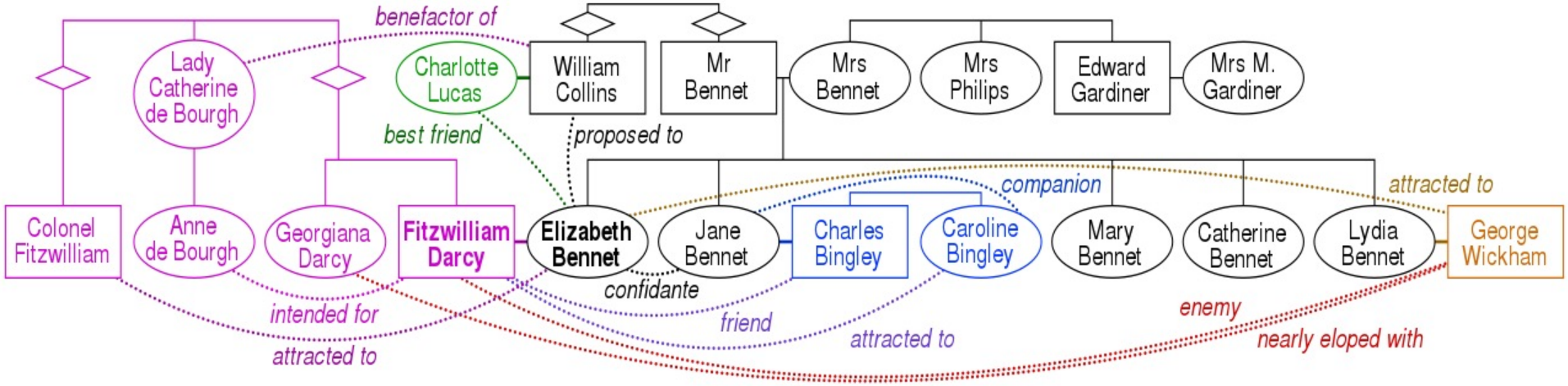
Natural Language Processing

10: Information Extraction – Part 2

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2021-06-10

Technology
Arts Sciences
TH Köln



parent(Mr. Bennet, Jane Bennet)

Information extraction

- Named entity recognition
- Relation extraction
- Entity linking

Named entity recognition

[tim cook]$_{PER}$ is the ceo of [apple]$_{ORG}$

- Identifying spans of text that correspond to typed entities

Main entities according to ACE (Automatic Content Extraction):

- Person (PER)
- Organization (ORG)
- Geo-political Entity (GPE)
- Location (LOC)
- Facility (FAC)
- Vehicle (VEH)
- Weapon (WEA)

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf

Relation extraction

The Big Sleep (1946 film)

From Wikipedia, the free encyclopedia

The Big Sleep is a 1946 American film noir directed by Howard Hawks,[2][3] the first film version of the 1939 novel of the same name by Raymond Chandler. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results".[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay. In 1997, the U.S. Library of Congress deemed the film "culturally, historically, or aesthetically significant," and added it to the National Film Registry.[5][6]

subject	predicate	object
The Big Sleep	directed_by	Howard Hawks
The Big Sleep	stars	Humphrey Bogart
The Big Sleep	stars	Lauren Bacall

Lehmann et al. (2012): DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia

Hearst patterns

pattern	sentence
NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, and other important civic buildings
NP_H such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP_H as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP_H {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP_H {,} especially {NP,}* {(or|and)} NP	European countries, especially France, England, and Spain

NP_H is the parent or hyponym

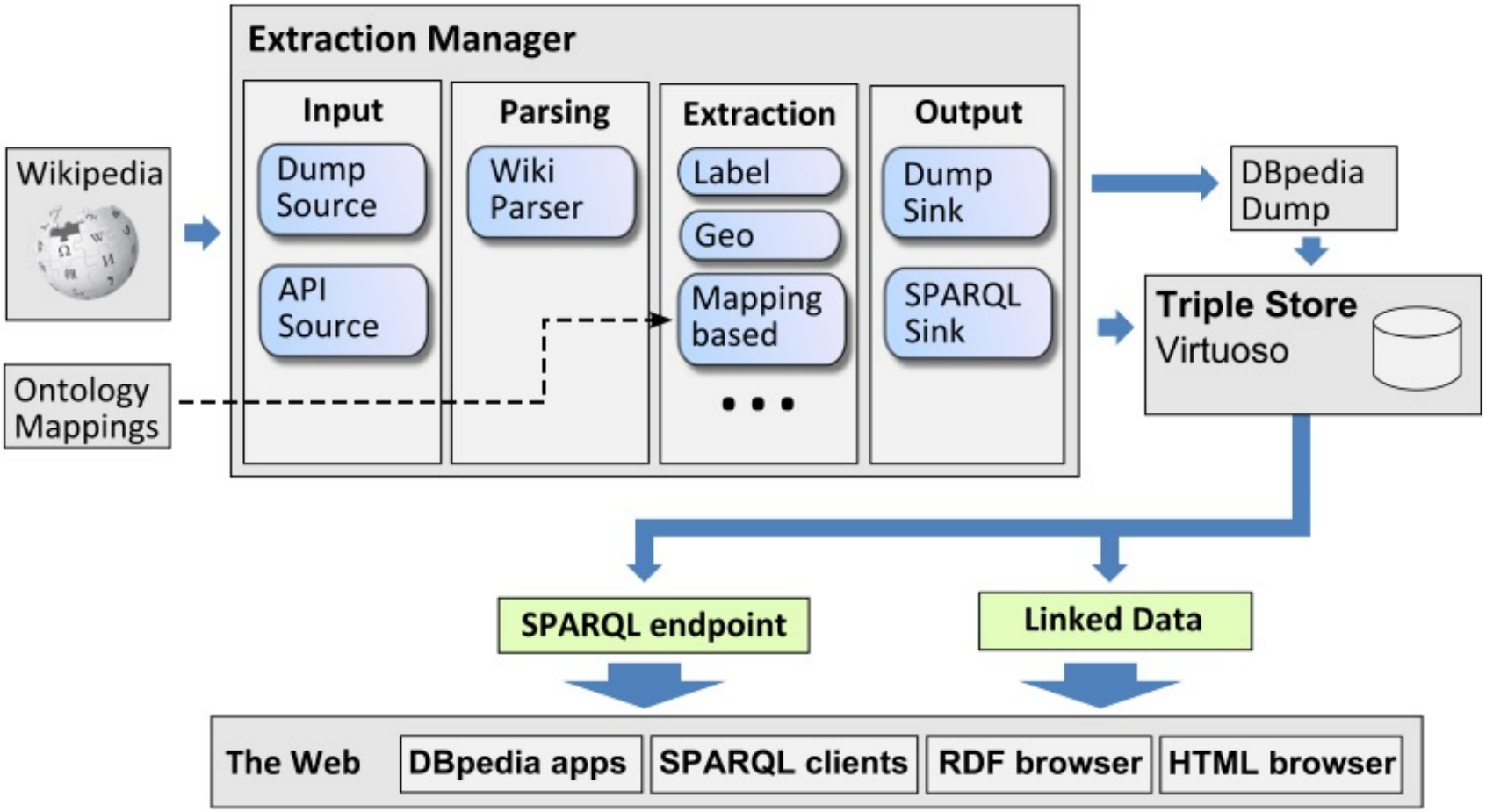
Hearst (1992): Automatic Acquisition of Hyponyms from Large Text Corpora - https://www.aclweb.org/anthology/C92-2082.pdf

Distant supervision

Start with seed relation: chancelor(Angela Merkel, Germany) and search in a large dataset (like the Web or Wikipedia)

- **Angela Merkel** (née Kasner; born 17 July 1954) is a **German** politician serving as the **chancellor** of **Germany** since 2005.
- **Chancellor Angela Merkel**, who hosted the online meeting, pledged further ...
- Biography of **German** politician **Angela Merkel**, who in 2005 became the first female **chancellor** of **Germany**.
- Annalena Baerbock, Greens ... The only woman in the race to succeed **Angela Merkel**, she is the Greens' first ever candidate for **chancellor**, as ...

DBpedia

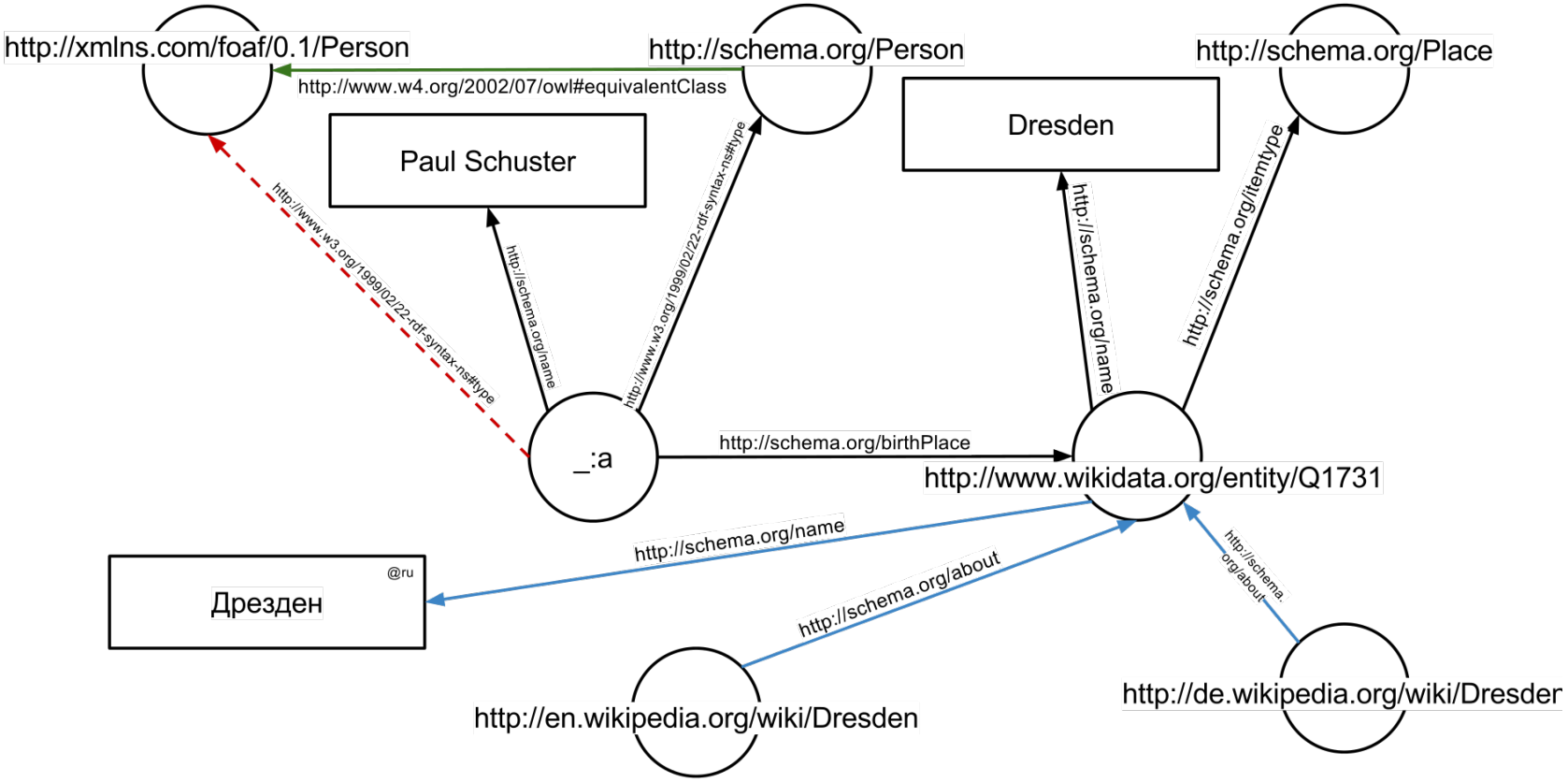


Lehmann et al. (2012): DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia

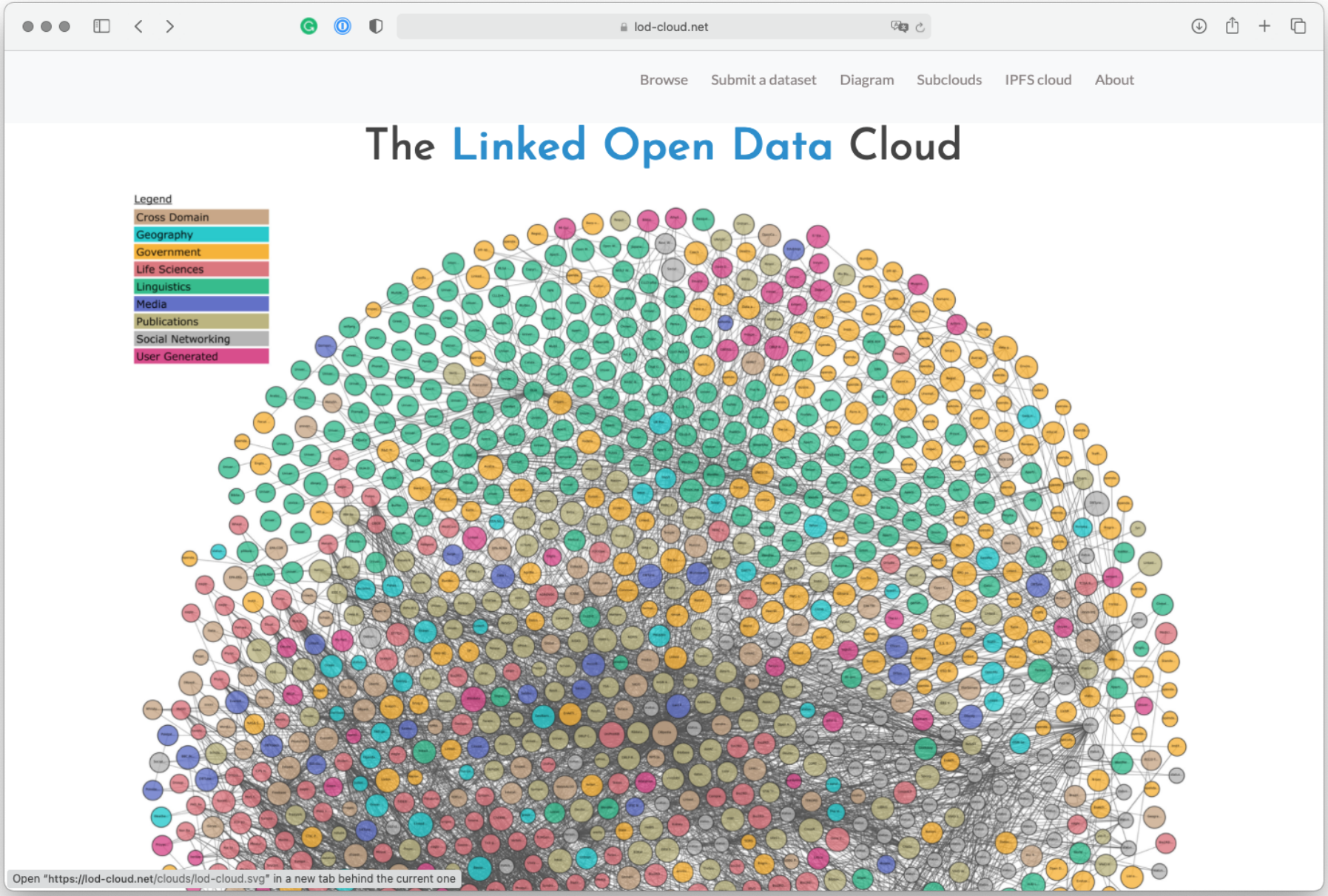
RDF – Resource Description Framework

```
<div vocab="http://schema.org/" typeof="Person">
    <span property="name">Paul Schuster</span> wurde in
    <span property="birthPlace" typeof="Place
    href="http://www.wikidata.org/entity/Q1731">
        <span property="name">Dresden</span>
    </span> geboren.
</div>
```

Linked (Open) Data



https://de.wikipedia.org/wiki/Semantic_Web#/media/Datei:RDF_example_extended.svg – CC0 by Denny Vrandečić



https://lod-cloud.net

Fun fact

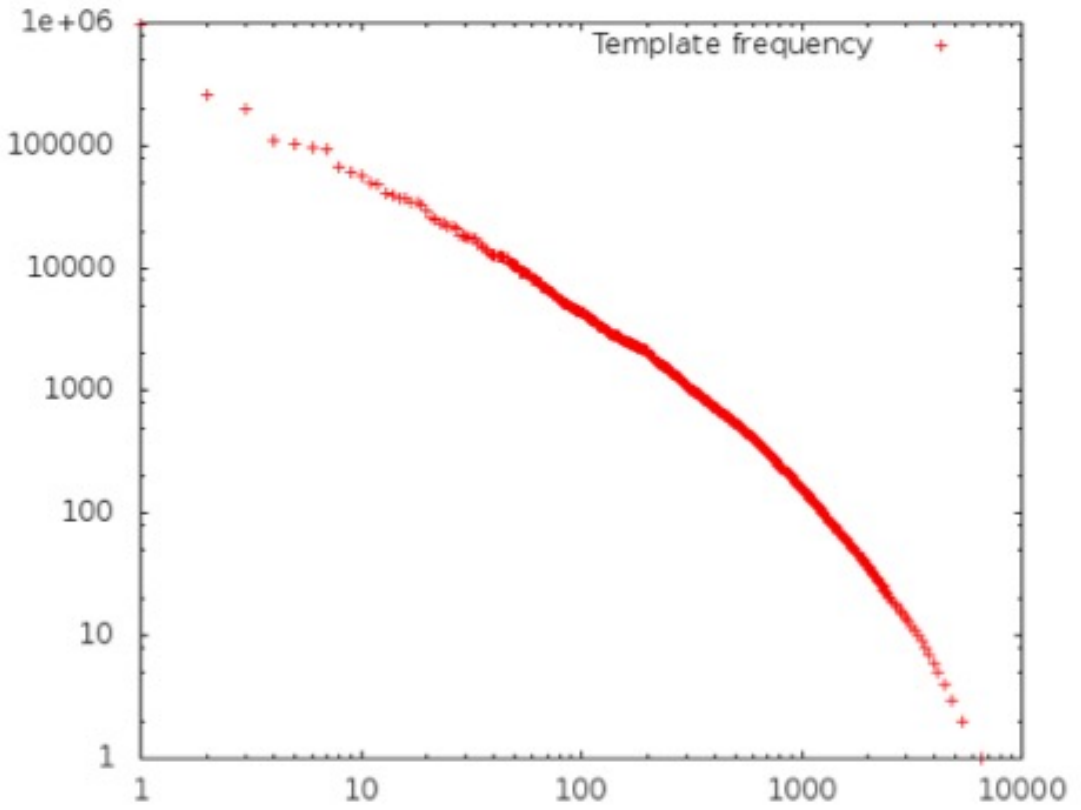
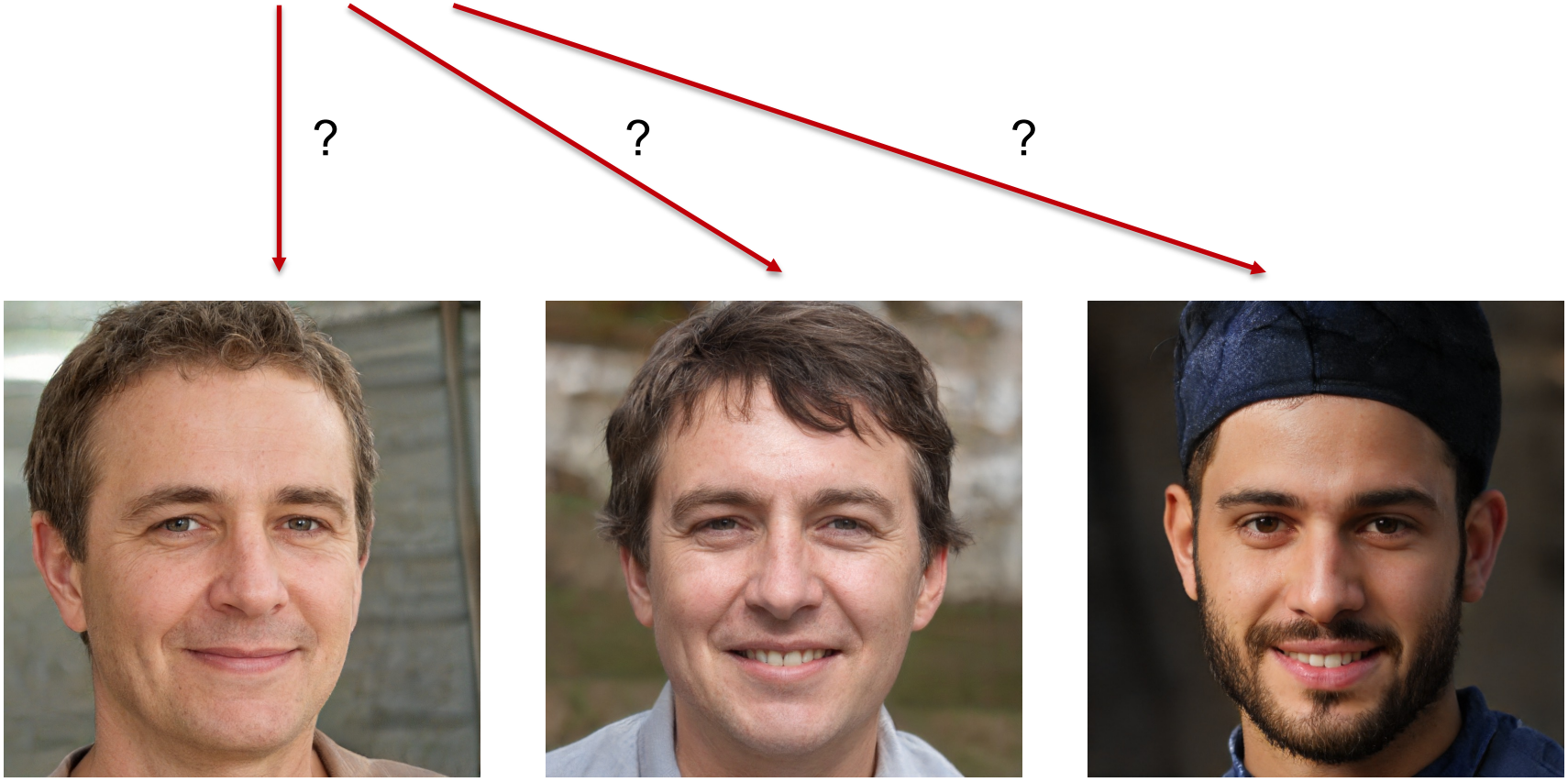


Fig. 7. English property mappings occurrence frequency (both axes are in log scale)

Lehmann et al. (2012): DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia

Entitiy Linking

- The Fluxcompensator engine was discovered in 2023 be "Michael de Blaunche"



https://thispersondoesnotexist.com – Karras et al. (2019)

Entity Linking aka Wikification

- Entity linking is the task of associating a mention in text with the representation of some real-world entity in an ontology

- Wikipedia is often the source of the text that answers the question. Each unique Wikipedia page acts as the unique id for a particular entity. This task of deciding which Wikipedia page corresponding to an individual is being referred to by a text mention has its own name: Wikification.

- Different approaches:
  - Use of anchor texts (TAGME): https://tagme.d4science.org/tagme/
  - Neural Graph-based linking

Ferragina, P. and Scaiella, U. (2011). Fast and accurate annotation of short texts with wikipedia pages

Fresh from the Research Front

TREC News Track

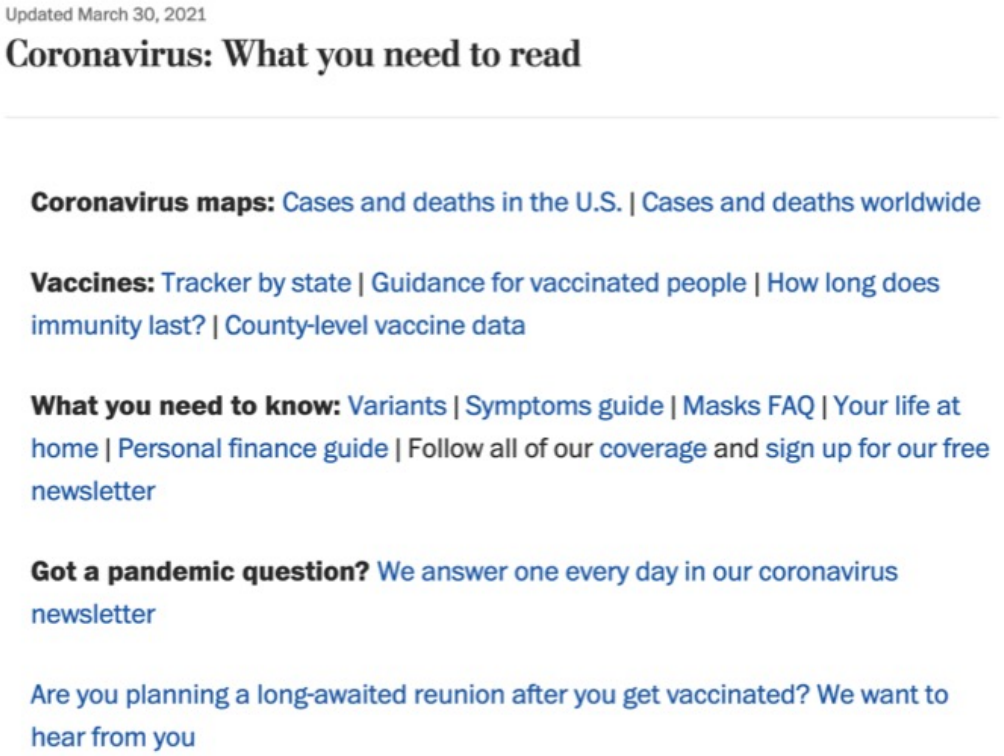
Task 1: Background linking

- Given a news story, retrieve other news articles that provide important context or background information. Build these links into an "explainer" box where links are grouped by what kind of context they offer. The goal is to help a reader understand or learn more about the story or main issues in the current article using the best possible sources.

Task 2: Wikification

- Given a news story, identify short passages in the article that should be hyperlinked to either another article, or a Wikipedia article, in order to provide in-context access to information that would help contextualize or fill in background on the story being read.

Background linkling



Updated March 30, 2021

Coronavirus: What you need to read

Coronavirus maps: Cases and deaths in the U.S. | Cases and deaths worldwide

Vaccines: Tracker by state | Guidance for vaccinated people | How long does immunity last? | County-level vaccine data

What you need to know: Variants | Symptoms guide | Masks FAQ | Your life at home | Personal finance guide | Follow all of our coverage and sign up for our free newsletter

Got a pandemic question? We answer one every day in our coronavirus newsletter

Are you planning a long-awaited reunion after you get vaccinated? We want to hear from you

The Washington Post implements manually linked "explainer boxes". This explainer box appeared at the end of most articles about COVID during the pandemic.

http://trec-news.org/guidelines-2021.pdf

TREC News Track

Task 1: Background linking

- Corpus with 728,626 newspaper articles from the Washington Post
- 51 Topics consisting of: Article, Description, Narrative

```
<top>
<num> Number: xxx </num>
<docid> f30b7db4-cc51-11e6-a747-d03044780a02</docid>
<url> https://www.washingtonpost.com/local/public-safety/homicides-remain-st eady-in-the-washington-region/2016/12/31/f30b7db4-cc51-11e6-a747-d03044780a 02_story.html</url>
<title> Topic title </title>
<desc> I would like to learn more about this topic </desc>
<narr> A traditional TREC narrative paragraph on the topic </narr>
<subtopics> <sub num="1">This is the first subtopic.</sub> <sub num="2">And this is the second one.</sub> </subtopics>
</top>
```

http://trec-news.org/guidelines-2021.pdf

Judging Relevance of Background Links

The relevance scale used by the NIST assessors was:

- 0: The linked document provides little or no useful background information.
- 1: The linked document provides some useful background or contextual information that would help the user understand the broader story context of the query article.
- 2: The document provides significantly useful background . . .
- 3: The document provides essential useful background . . .
- 4: The document MUST appear in the sidebar otherwise critical context is missing.

TREC News Track

Example

Query article: Love in the time of climate change: Grizzlies and polar bears are now mating (May 23, 2016)

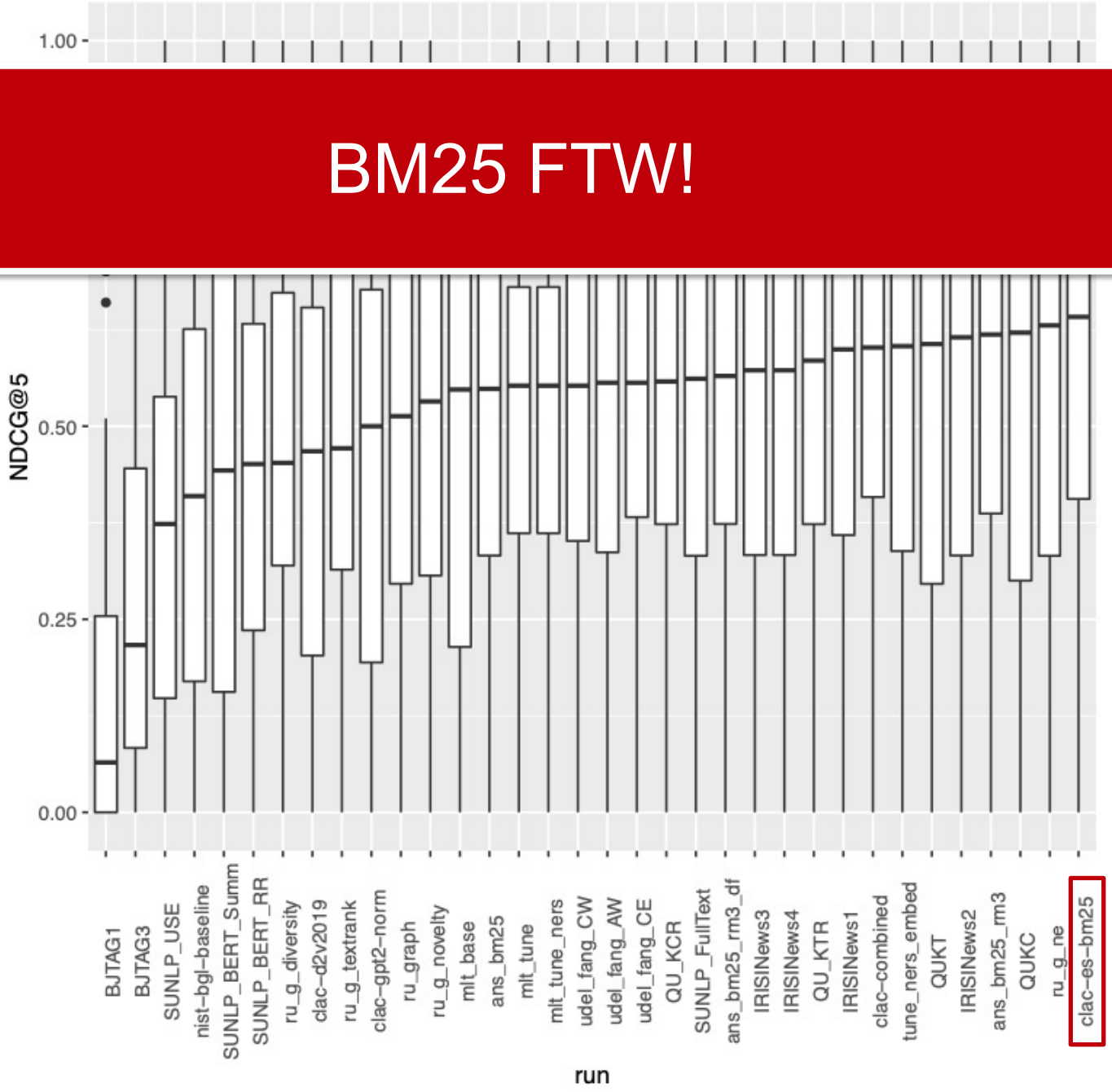
This article describes and analyzes a phenomenon where grizzlies and polar bears are mating to create a new species known as pizzlies or grolars. It explains why this is happening and points out that it happens (or has happened) to other species as well. Articles along these lines are good background links. For example:

- Coywolves, coyote-wolf hybrids, are prowling Rock Creek Park and D.C. suburbs (July 1, 2014)
- Humans and Neanderthals may have interbred 50,000 years earlier than previously thought (February 17, 2016)

However, the following article is of less relevance and should be ranked lower because it's not about interbreeding.

- Why do seals keep trying to have sex with penguins? (November 18, 2014)

BM25 FTW!


1.00
0.50
0.25
0.00
NDCG@5
BJTAG1
BJTAG3
SUNLP_USE
nist-bgl-baseline
SUNLP_BERT_Summ
SUNLP_BERT_RR
ru_g_diversity
clac-d2v2019
ru_g_textrank
clac-gpt2-norm
ru_graph
ru_g_novelty
mlt_base
ans_bm25
mlt_tune
mlt_tune_ners
udel_fang_CW
udel_fang_AW
udel_fang_CE
QU_KCR
SUNLP_FullText
ans_bm25_rm3_df
IRISINews3
IRISINews4
QU_KTR
IRISINews1
clac-combined
tune_ners_embed
QUKT
IRISINews2
ans_bm25_rm3
QUKC
ru_g_ne
clac-es-bm25
run

Our approach

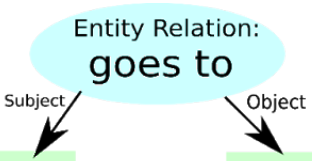
- Large corpus makes costly NLP techniques for all documents infeasible.
  - Use index and TF-IDF to select (approx. 200 documents/topic) interesting articles.
  - Extract entities and relations for selected documents.

- Assumption: Articles are more likely to be relevant if both articles share many common relations.

Thx @Björn

Our approach

- Entities often occur by chance, possibly in different contexts, in both articles:
  - **Nixon** *goes to* **China.**
  - **Nixon** was born in 1913. … In 2010, China became the world's second largest economy.



The shorthand is "Nixon goes to China," meaning a moment in which a leader reverses his past positions to do something that is shocking but beneficial.

Thx @Björn

Our approach

Extract all entities for each of the selected documents (spacy):

- New York, Trump, Labor Department, Nasa, Air Force

Find relations between the entities:

- Nixon visits China, Obama pass healthcare

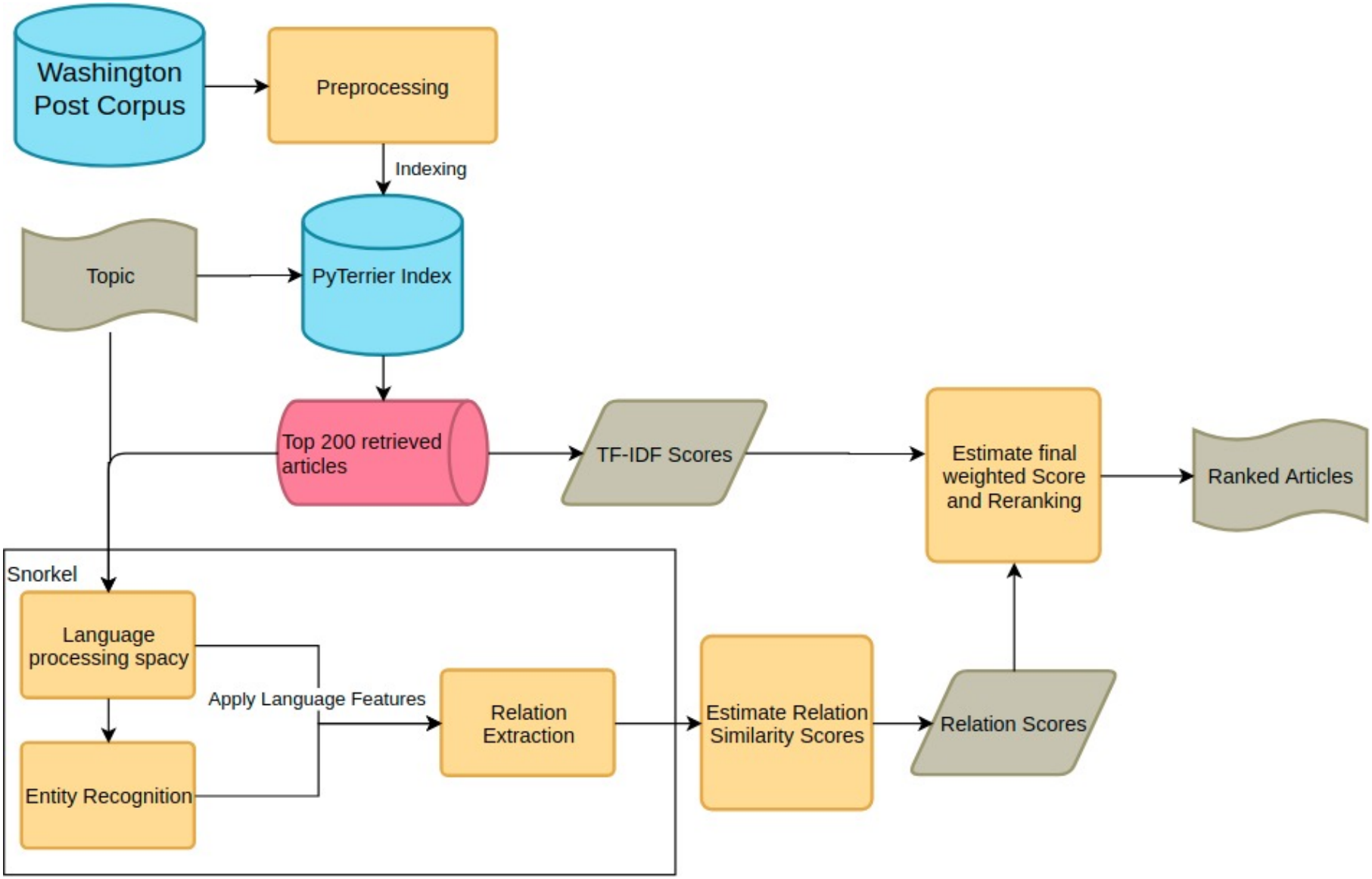
Simple recognition of relations via linguistic features (Snorkel):

- Do entities appear in the same sentence?
- Is there a subject-object relationship between entities?
- Which verb connects the entities?

Give more weight to rare entities (IDF).

Final score for document calculated from TF-IDF score and weighted amount of shared relations.

Thx @Björn


Washington Post Corpus
Preprocessing
Indexing
Topic
PyTerrier Index
Top 200 retrieved articles
TF-IDF Scores
Estimate final weighted Score and Reranking
Ranked Articles
Snorkel
Language processing spacy
Apply Language Features
Entity Recognition
Relation Extraction
Estimate Relation Similarity Scores
Relation Scores

Thx @Björn

Real-world Example

Retrieved Example:

- **Topic:** Olympics 2020
- **Topic Title:** For 2020 Olympic hopefuls, postponement is another challenge to overcome.
- **Title of retrieved Article:** For finely tuned Olympic athletes, a one-year postponement changes everything
- **Shared Relation Entities:**
  - 'U.S.' - 'Tokyo'
  - 'British' - 'Adam Peaty'
  - 'Helen Maroulis' - 'Olympic'
  - 'American' - 'Emma Coburn'

Thx @Björn

The Data Programming Paradigm

In many applications, we would like to use machine learning, but we face the following challenges:

- (i) *hand-labeled* training data is not available, and is prohibitively expensive to obtain in sufficient quantities (domain experts)
- (ii) *related external knowledge bases* are either unavailable or insufficiently specific, precluding a traditional distant supervision;
- (iii) *application specifications* are in flux, changing the model we ultimately wish to learn.

In data programming, rather than manually labeling each example, users instead describe the *processes by which* these points could be labeled by providing a set of heuristic rules called *labeling functions*.

https://arxiv.org/abs/1605.07723

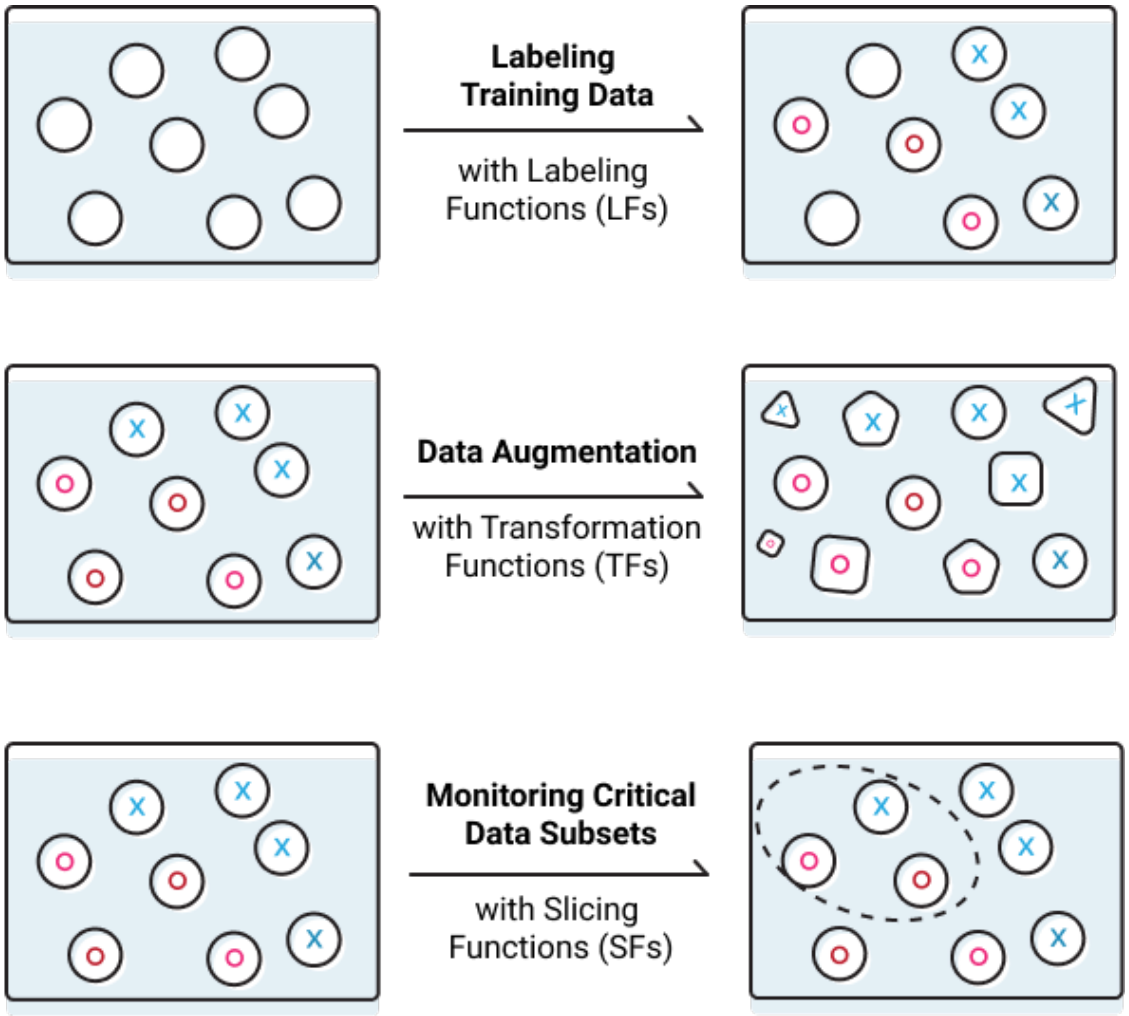
The Data Programming Paradigm

Labeling function

- need not have perfect accuracy or recall;
- rather, it represents a pattern that the user wishes to impart to their model and that is easier to encode as a labeling function than as a set of hand-labeled examples.
- labeling functions can be based on external knowledge bases, libraries or ontologies, can express heuristic patterns, or some hybrid of these types;
- labeling functions are more general than manual annotations, as a manual annotation can always be directly encoded by a labeling function;
- labeling functions can overlap, conflict, and even have dependencies which users can provide as part of the data programming specification

https://arxiv.org/abs/1605.07723

Introducing: Snorkel



https://www.snorkel.org/get-started/

Demo

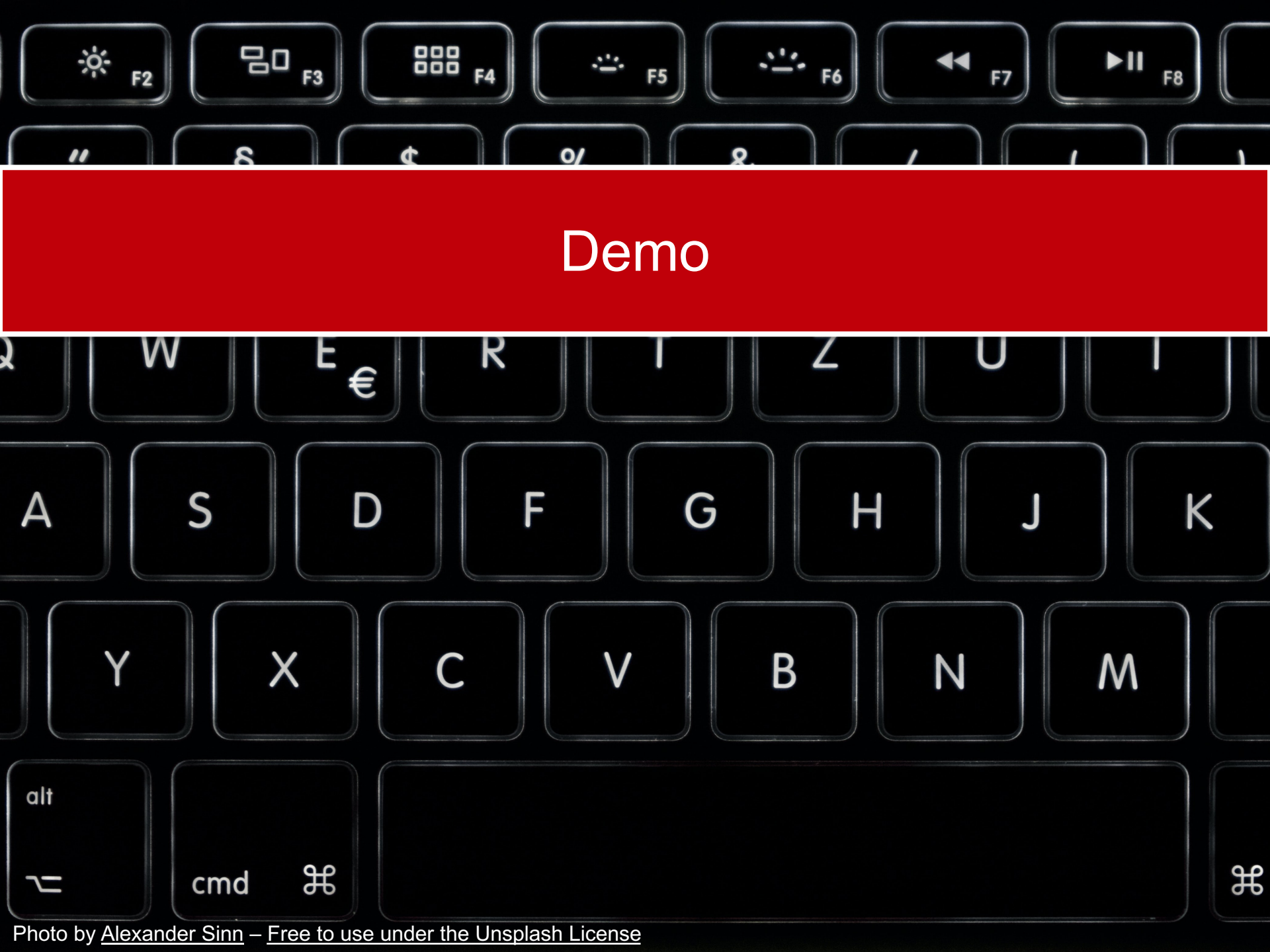
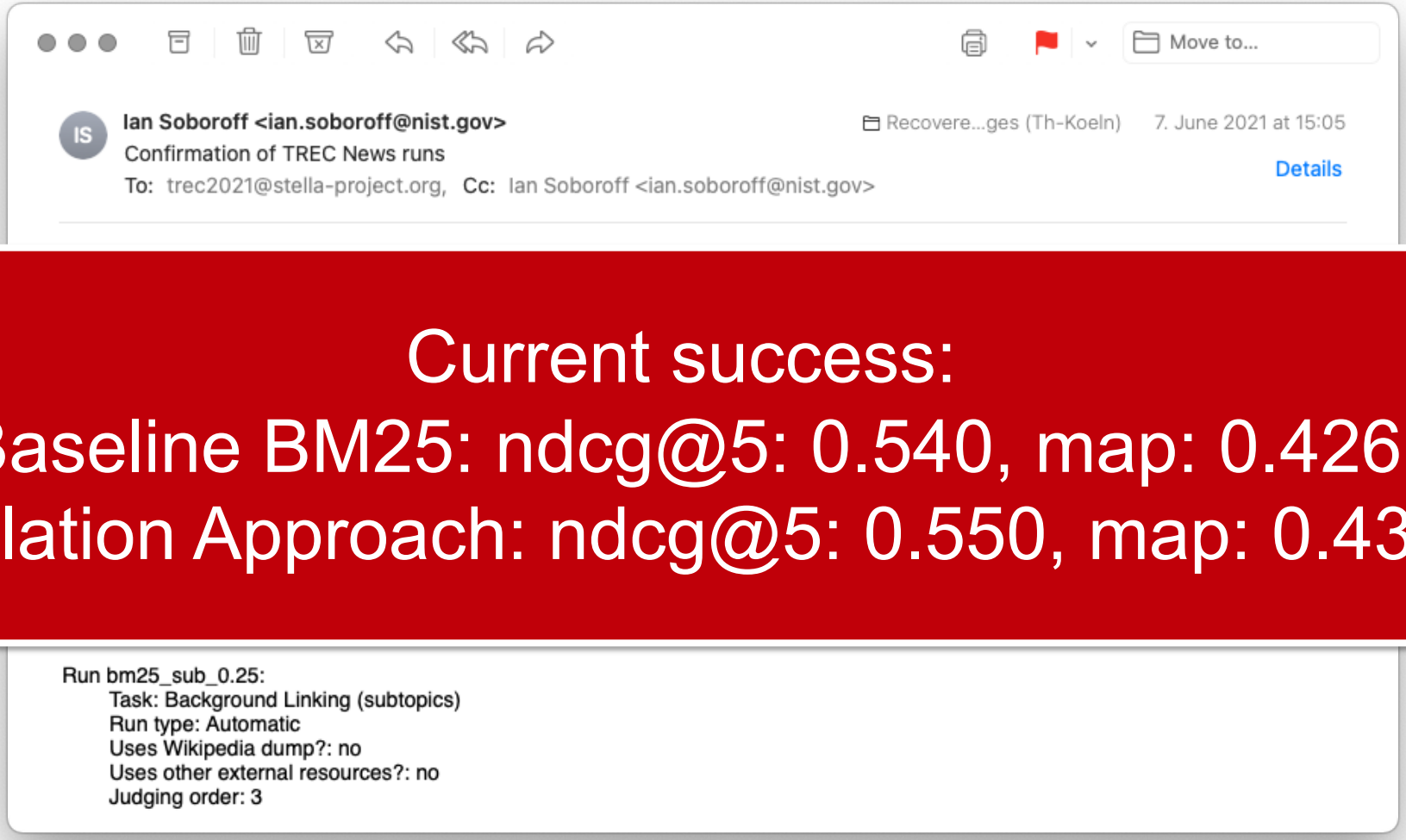


Photo by Alexander Sinn – Free to use under the Unsplash License

Let's see …



Current success:

Baseline BM25: ndcg@5: 0.540, map: 0.426

Relation Approach: ndcg@5: 0.550, map: 0.430