BBC Research & Development

Semantic Tagging and Text Classification for Archives and Content Production Systems

Chris Newell

- Lead R&D Engineer
- Natural Language Processing
- Internet Research and Future Services

15 June 2021

ISKO UK Meetup

SEMANTIC TAGGING AND TEXT CLASSIFICATION

BBC Research & Development

- **The BBC's Royal Charter mandates "a centre of excellence" for research and development in the electronic distribution of audio-visual media**
  - 200 research engineers, scientists and designers
  - Research Labs in London and Salford
- **Work on every aspect of broadcast and online services**
  - Production
  - Distribution
  - Audiences

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Natural Language Processing at BBC R&D

- Semantic Tagging
- Text Classification
- Speech to Text
- Text Summarisation
- Sentiment Analysis
- Quote Extraction

Applied to:

- Web content
- News feeds
- TV subtitles
- Radio transcripts
- Social media

SEMANTIC TAGGING AND TEXT CLASSIFICATION

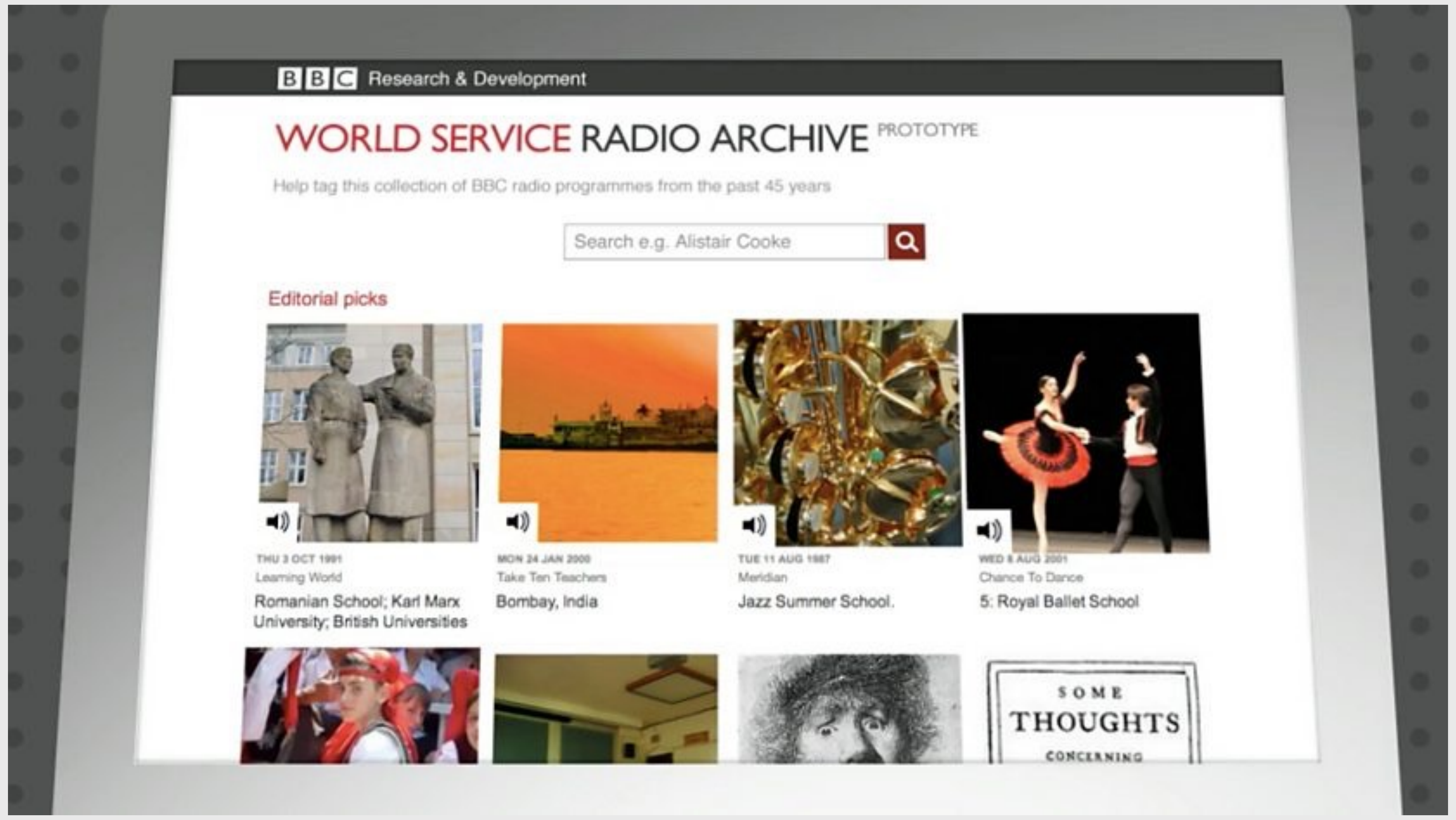
Semantic Tagging and Text Classification

- **Describing a document in terms of unambiguous entities and themes**
- **Linked to additional sources of information**
- **Tagging can consist of either**
    - all the entities and themes mentioned in the text
    - the primary topic(s) of the text

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Motivation for Semantic Tagging

- World Service Archive Project (2011- 2014)
- 70,000 radio programmes from over 60 years
- Metadata sparse, inaccurate or non-existent
- Efficient access and re-use of content is impossible without metadata



BBC Research & Development

WORLD SERVICE RADIO ARCHIVE PROTOTYPE

Help tag this collection of BBC radio programmes from the past 45 years

Search e.g. Alistair Cooke

Editorial picks

THU 3 OCT 1991
Learning World
Romanian School; Karl Marx University; British Universities

MON 24 JAN 2000
Take Ten Teachers
Bombay, India

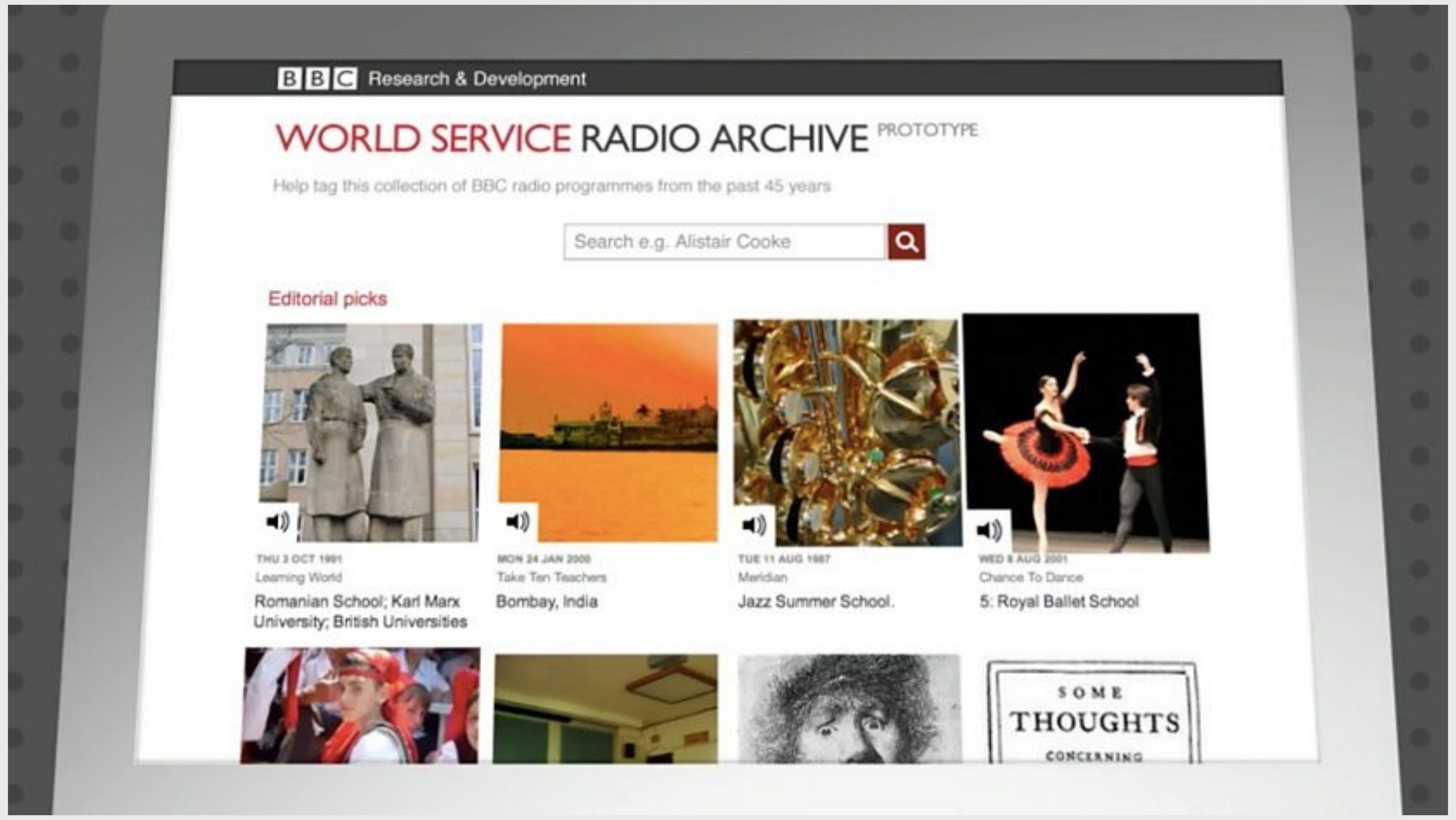
TUE 11 AUG 1987
Meridian
Jazz Summer School.

WED 8 AUG 2001
Chance To Dance
5: Royal Ballet School

SEMANTIC TAGGING AND TEXT CLASSIFICATION

World Service Archive Approach

- Speech-to-text technology used to create transcripts
- Tags extracted using an automated tagging system called Mango
- Tagging supported search and navigation in prototype
- Crowd sourcing used to improve the tagging



BBC Research & Development

WORLD SERVICE RADIO ARCHIVE PROTOTYPE

Help tag this collection of BBC radio programmes from the past 45 years

Search e.g. Alistair Cooke

Editorial picks

THU 3 OCT 1991
Learning World
Romanian School; Karl Marx University; British Universities

MON 24 JAN 2000
Take Ten Teachers
Bombay, India

TUE 11 AUG 1987
Meridian
Jazz Summer School.

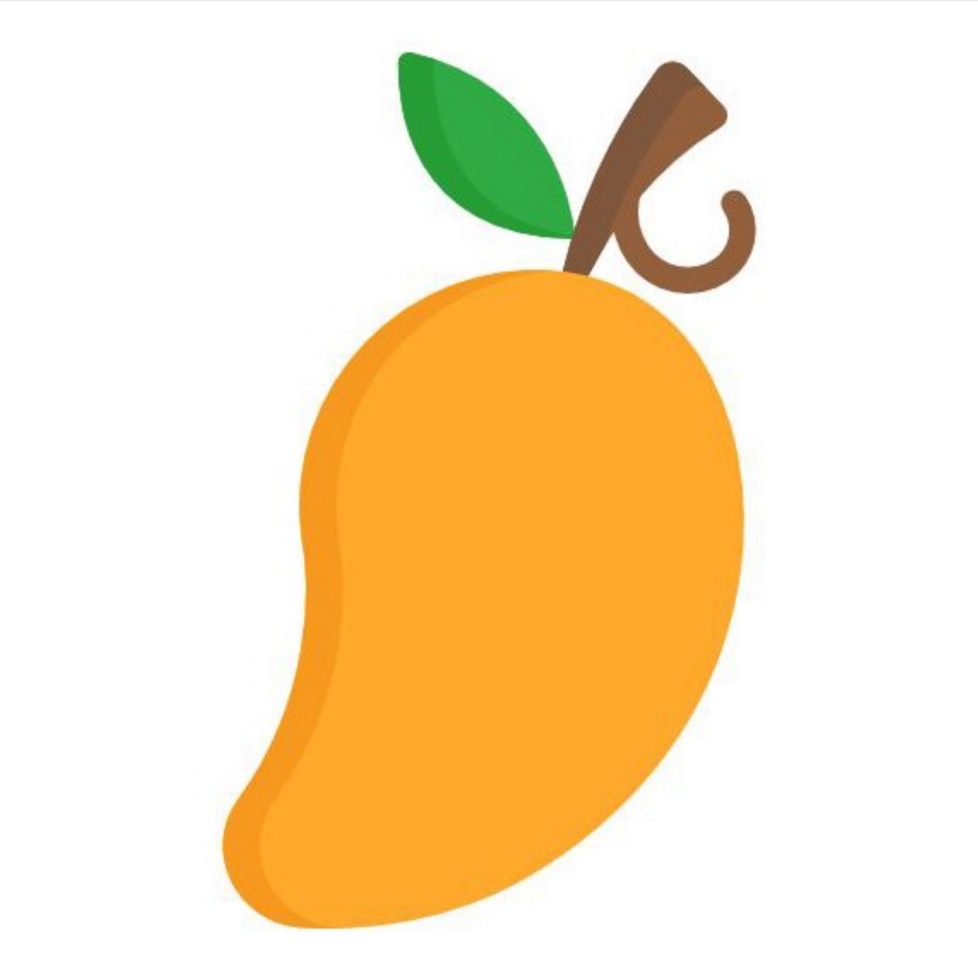
WED 8 AUG 2001
Chance To Dance
5: Royal Ballet School

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Mango Overview

- **Mango is a semantic tagging system that identifies entities and themes e.g. for the text**

  “The German chancellor, Angela Merkel called on the Russian president to exercise his influence on the separatists to enforce the ceasefire”

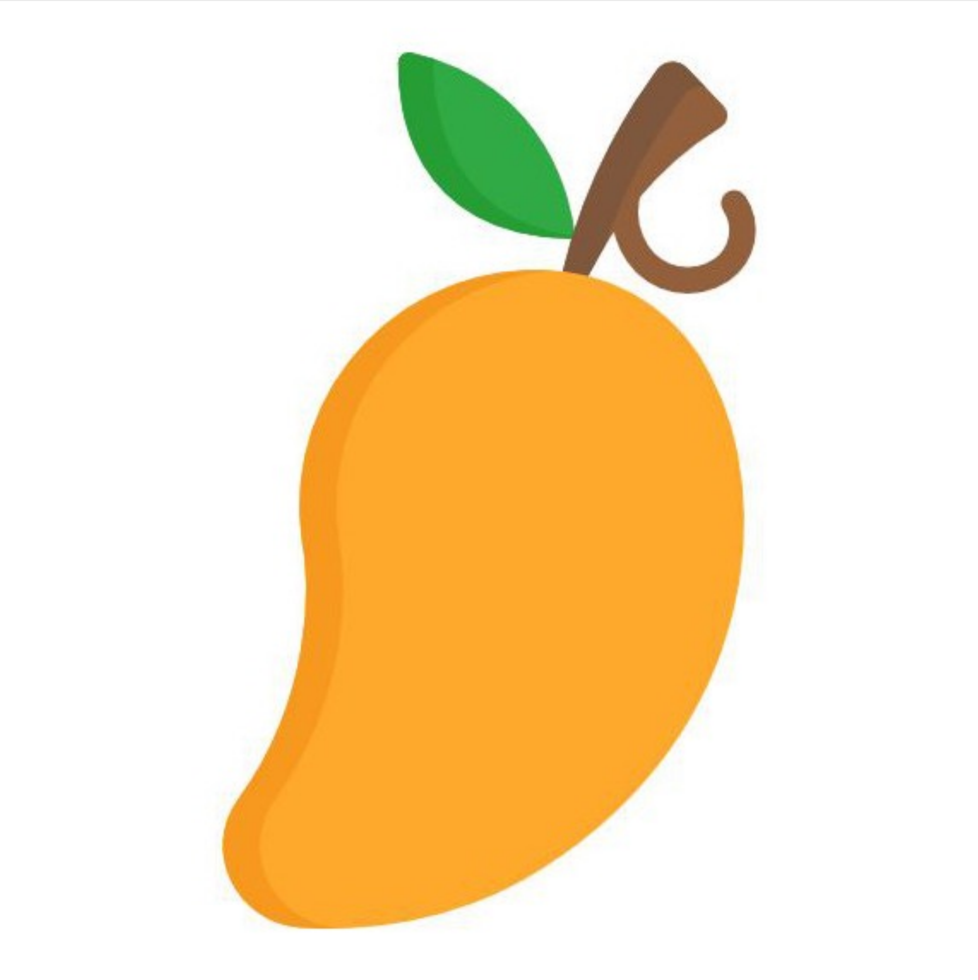


SEMANTIC TAGGING AND TEXT CLASSIFICATION

Mango Overview

- **Mango is a semantic tagging system that identifies entities and themes e.g. for the text**

  “The German chancellor, Angela Merkel called on the Russian president to exercise his influence on the separatists to enforce the ceasefire”



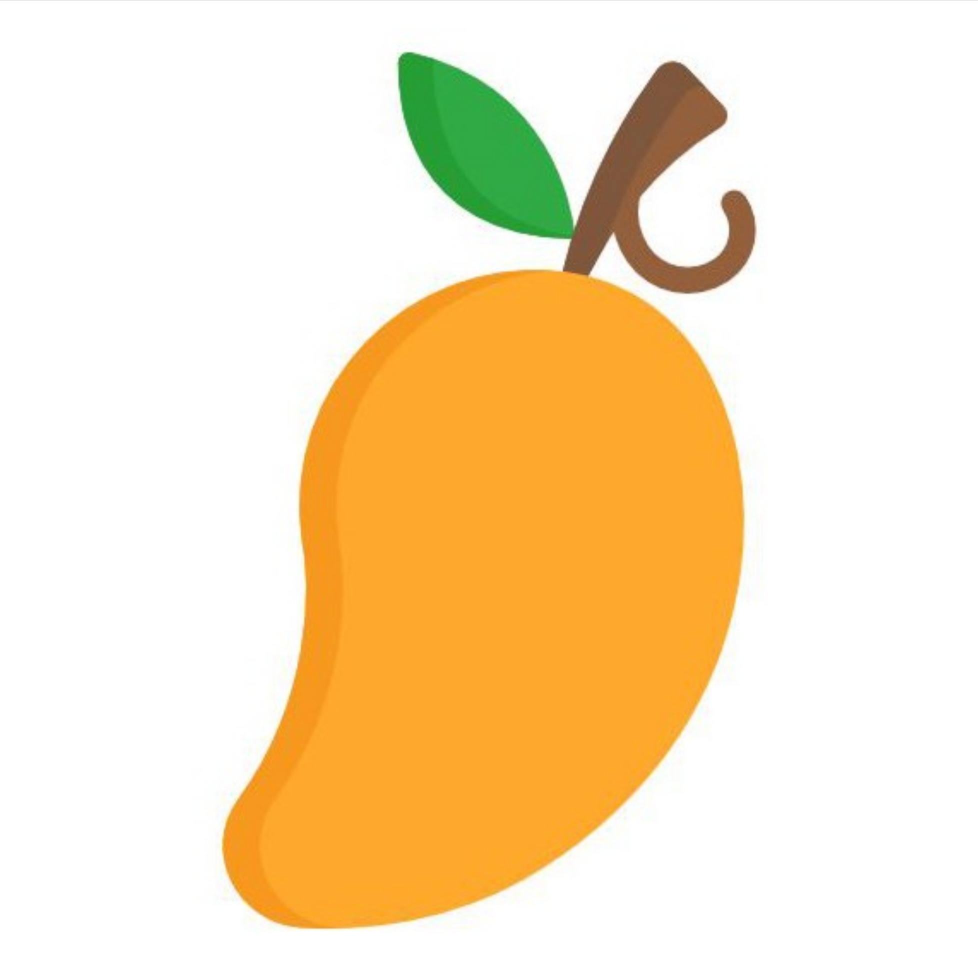
SEMANTIC TAGGING AND TEXT CLASSIFICATION

Mango Overview

- **Mango is a semantic tagging system that identifies entities and themes e.g. we want to identify:**

  "The German chancellor, Angela Merkel called on the Russian president to exercise his influence on the separatists to enforce the ceasefire"

- **Mango uses DBpedia identifiers:**
  - http://dbpedia.org/resource/Angela_Merkel
  - http://dbpedia.org/resource/Chancellor_of_Germany
  - http://dbpedia.org/resource/President_of_Russia



SEMANTIC TAGGING AND TEXT CLASSIFICATION

DBpedia entry for Angela_Merkel

dbpedia-owl:birthDate	1954-07-17
dbpedia-owl:birthPlace	dbpedia:West_Germany dbpedia:Hamburg
dbpedia-owl:alias	Merkel, Angela Dorothea (full name)
dbpedia-owl:almaMater	dbpedia:Leipzig_University
dbpedia-owl:office	Minister of the Environment Chancellor of Germany Leader of the Christian Democratic Union Member of the Bundestag Minister of Women and Youth
dbpedia-owl:party	dbpedia:Christian_Democratic_Union

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Mango Components

- **Mango has two components:**
  - Spotter
  - Disambiguator
- **Both created using a complete dump of Wikipedia**
  - https://dumps.wikimedia.org/enwiki

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Spotter Index

- **The spotter uses an index which maps surface forms to entities**
- **For example:**
  - “Birmingham City F.C.”
  - “Birmingham City”
  - “Birmingham”
  - “The Blues”
- **All map to:**
  - http://dbpedia.org/resource/Birmingham_City_F.C.

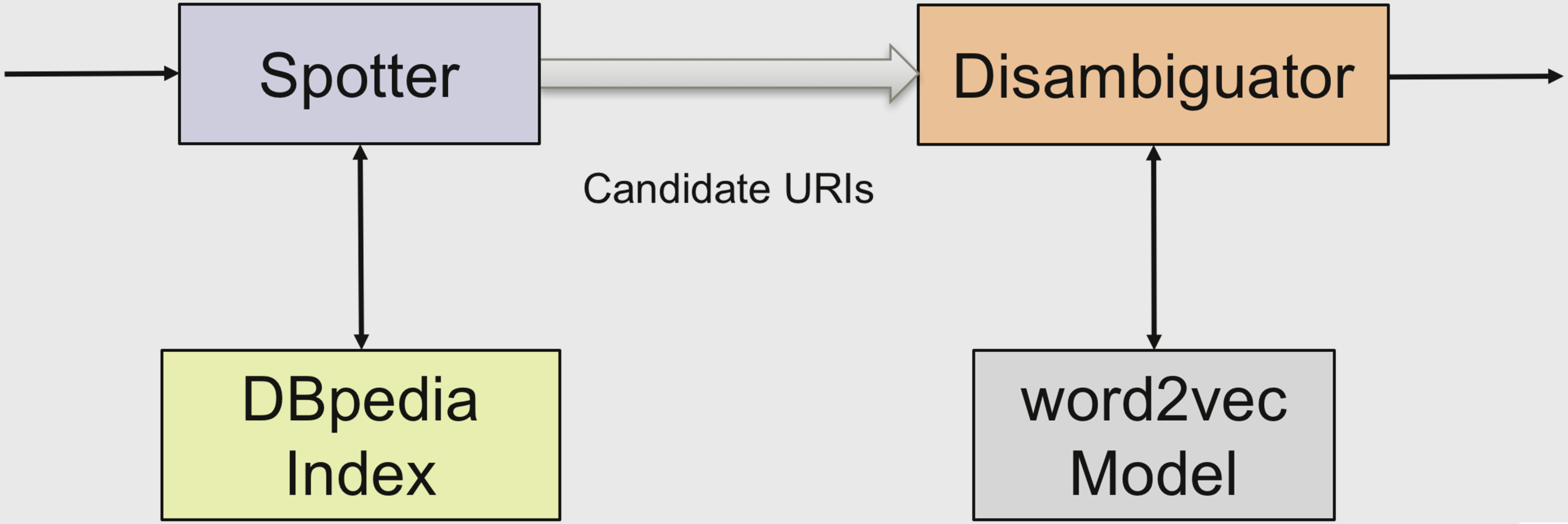
SEMANTIC TAGGING AND TEXT CLASSIFICATION

Ambiguity Problem

- **However many surface forms are ambiguous**
- **“Birmingham” could refer to either:**
  - http://dbpedia.org/resource/Birmingham_City_F.C.
  - http://dbpedia.org/resource/Birmingham
  - http://dbpedia.org/resource/Birmingham,_Alabama
  - http://dbpedia.org/resource/Birmingham_(horse)
  - etc…

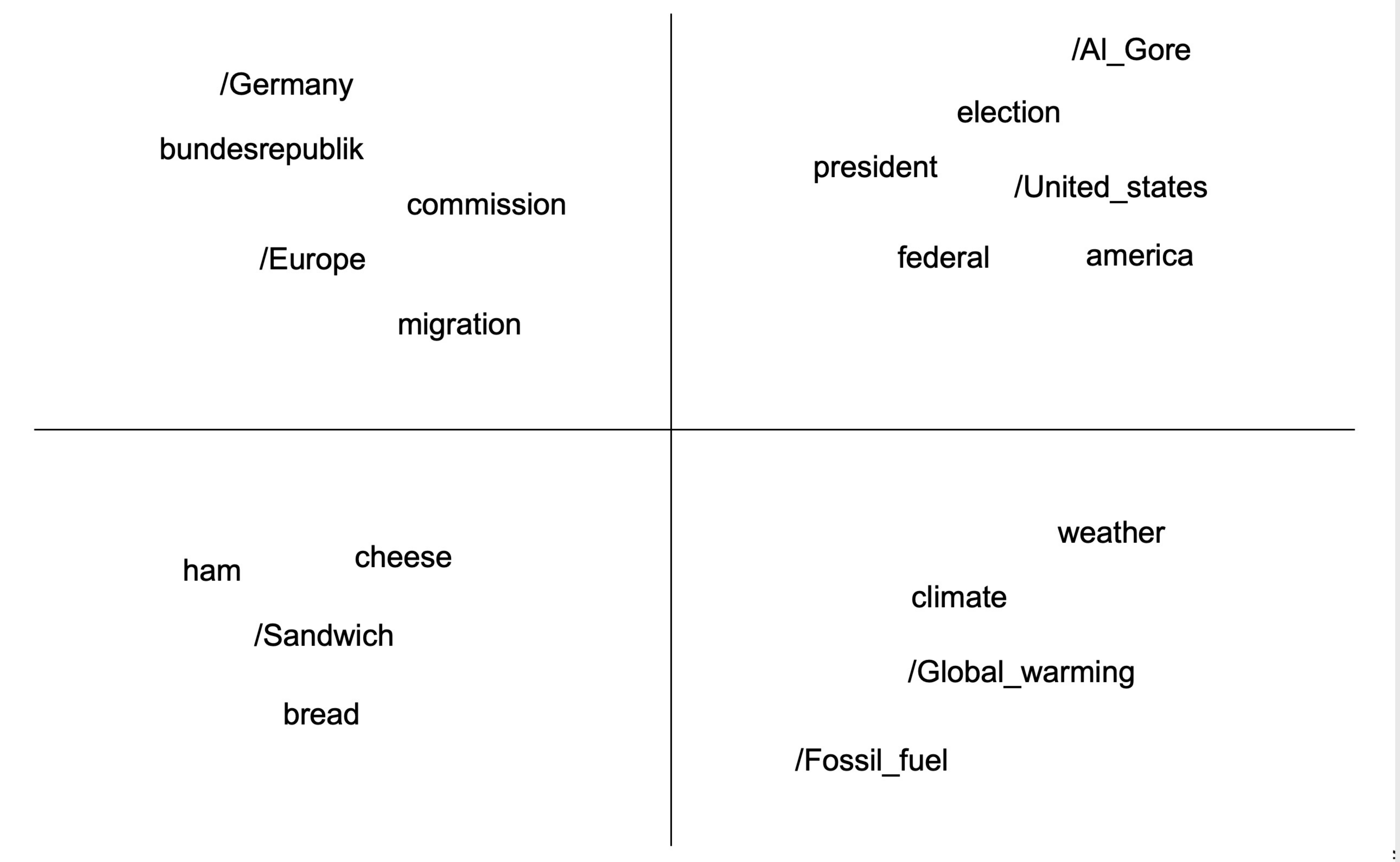
SEMANTIC TAGGING AND TEXT CLASSIFICATION

Mango Schematic



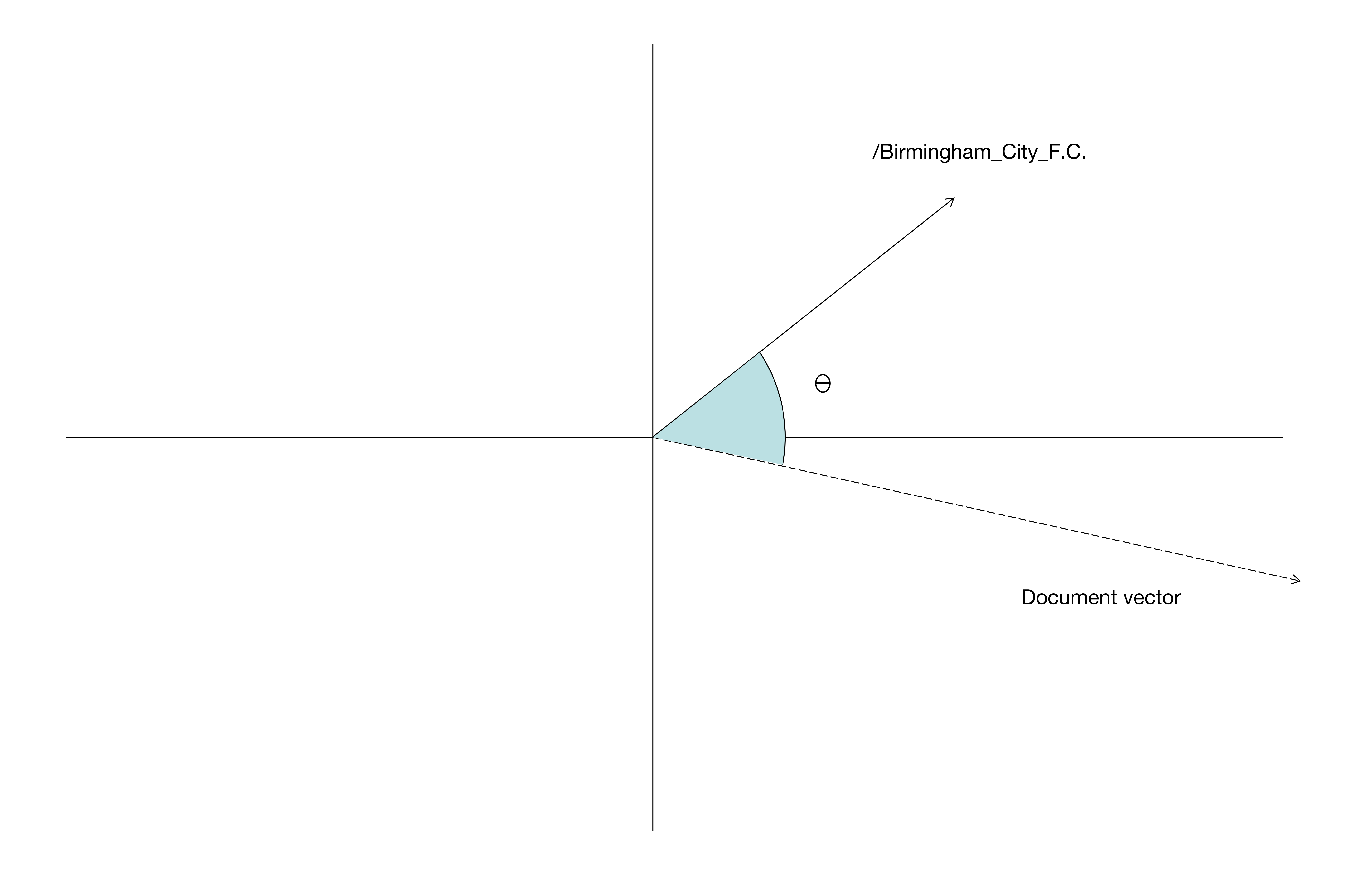
SEMANTIC TAGGING AND TEXT CLASSIFICATION

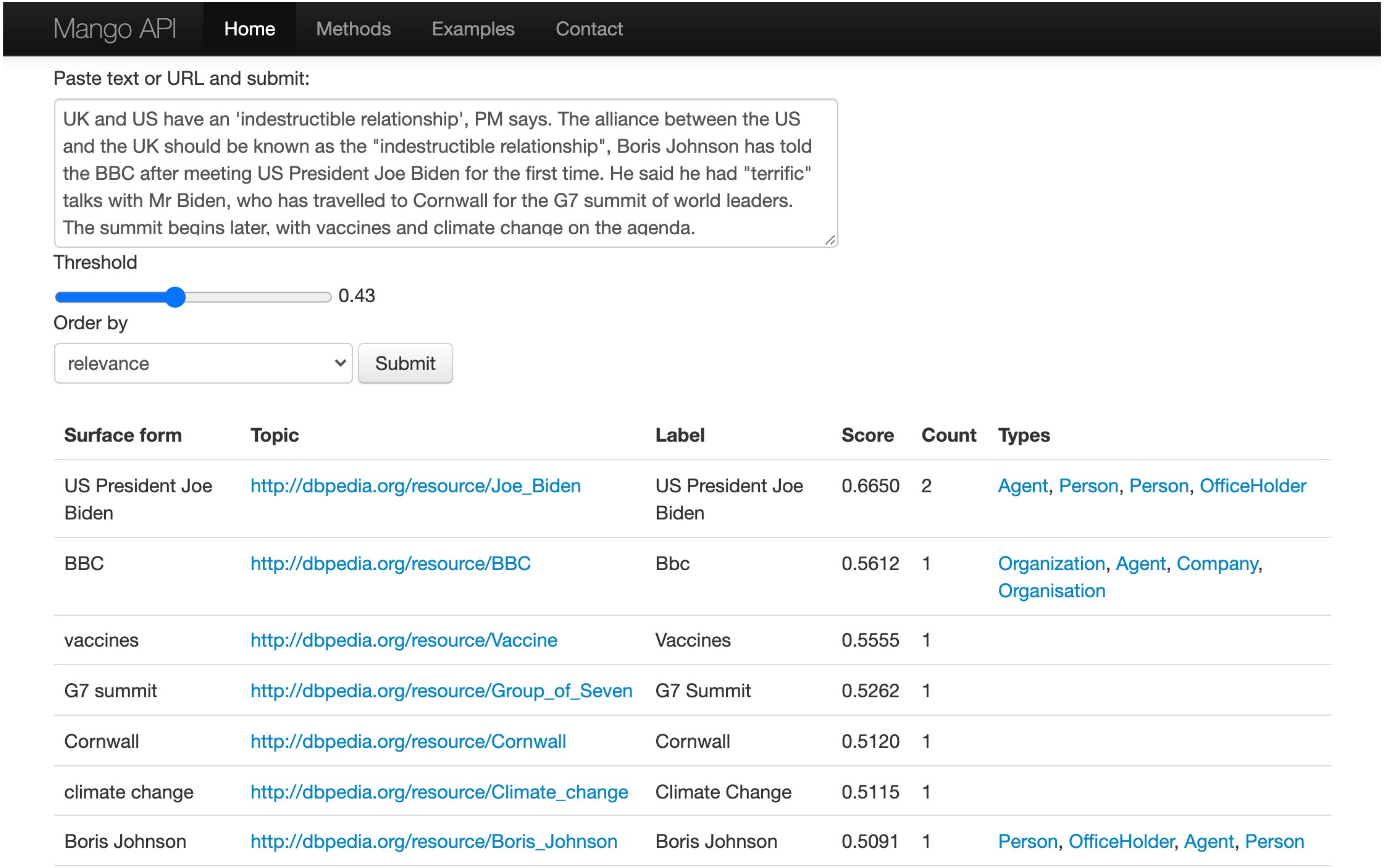
Word2Vec:
A vector space representation of words



SEMANTIC TAGGING AND TEXT CLASSIFICATION

Comparing Candidates against Documents





Paste text or URL and submit:

UK and US have an 'indestructible relationship', PM says. The alliance between the US and the UK should be known as the "indestructible relationship", Boris Johnson has told the BBC after meeting US President Joe Biden for the first time. He said he had "terrific" talks with Mr Biden, who has travelled to Cornwall for the G7 summit of world leaders. The summit begins later, with vaccines and climate change on the agenda.

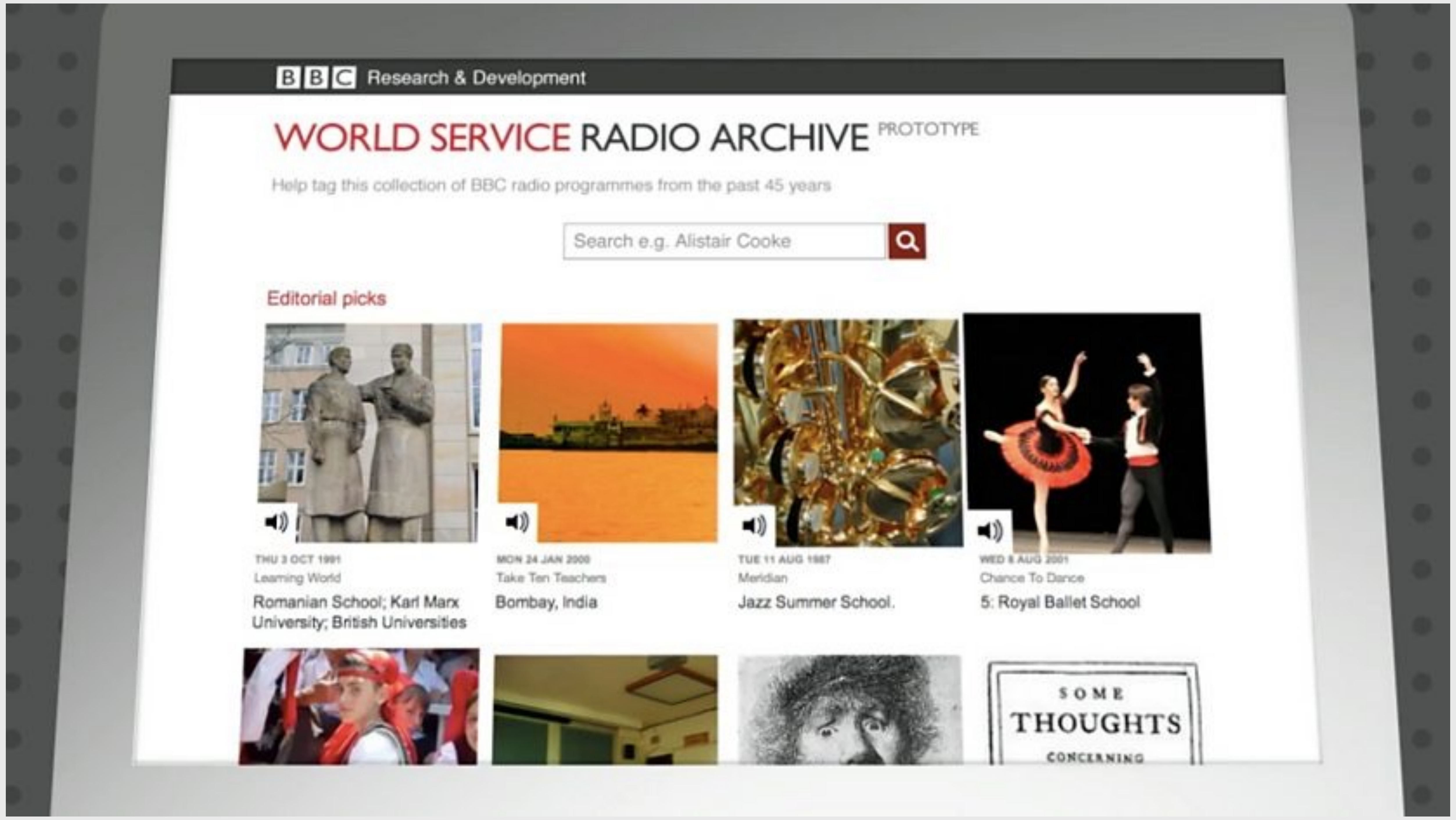
Threshold 0.43

Order by relevance

Surface form	Topic	Label	Score	Count	Types
US President Joe Biden	http://dbpedia.org/resource/Joe_Biden	US President Joe Biden	0.6650	2	Agent, Person, Person, OfficeHolder
BBC	http://dbpedia.org/resource/BBC	Bbc	0.5612	1	Organization, Agent, Company, Organisation
vaccines	http://dbpedia.org/resource/Vaccine	Vaccines	0.5555	1	
G7 summit	http://dbpedia.org/resource/Group_of_Seven	G7 Summit	0.5262	1	
Cornwall	http://dbpedia.org/resource/Cornwall	Cornwall	0.5120	1	
climate change	http://dbpedia.org/resource/Climate_change	Climate Change	0.5115	1	
Boris Johnson	http://dbpedia.org/resource/Boris_Johnson	Boris Johnson	0.5091	1	Person, OfficeHolder, Agent, Person

SEMANTIC TAGGING AND TEXT CLASSIFICATION

World Service Archive



BBC Research & Development

WORLD SERVICE RADIO ARCHIVE PROTOTYPE

Help tag this collection of BBC radio programmes from the past 45 years

Search e.g. Alistair Cooke

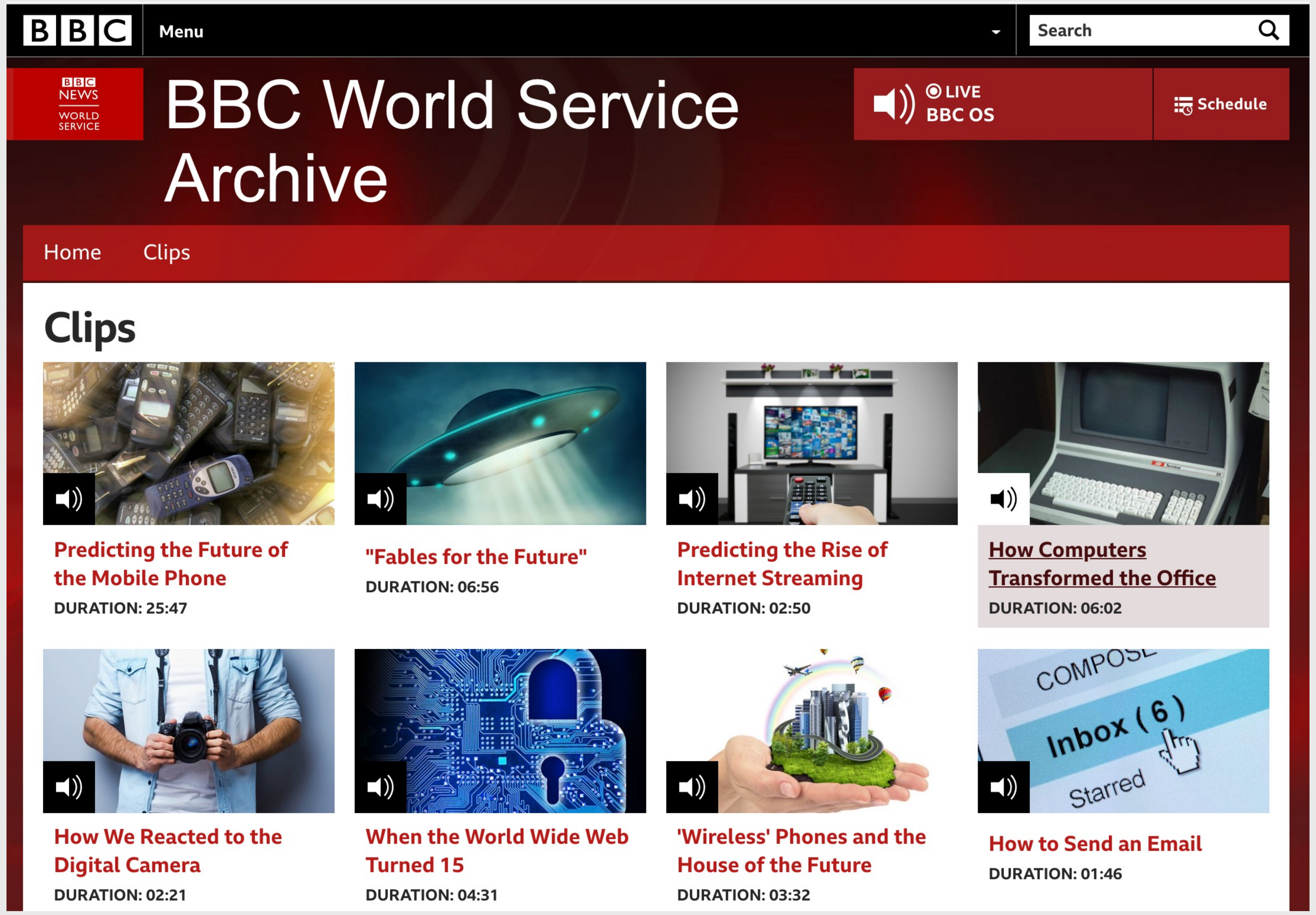
Editorial picks

THU 3 OCT 1991
Learning World
Romanian School; Karl Marx University; British Universities

MON 24 JAN 2000
Take Ten Teachers
Bombay, India

TUE 11 AUG 1987
Meridian
Jazz Summer School.

WED 8 AUG 2001
Chance To Dance
5: Royal Ballet School



BBC World Service Archive

Clips

Predicting the Future of the Mobile Phone
DURATION: 25:47

"Fables for the Future"
DURATION: 06:56

Predicting the Rise of Internet Streaming
DURATION: 02:50

How Computers Transformed the Office
DURATION: 06:02

How We Reacted to the Digital Camera
DURATION: 02:21

When the World Wide Web Turned 15
DURATION: 04:31

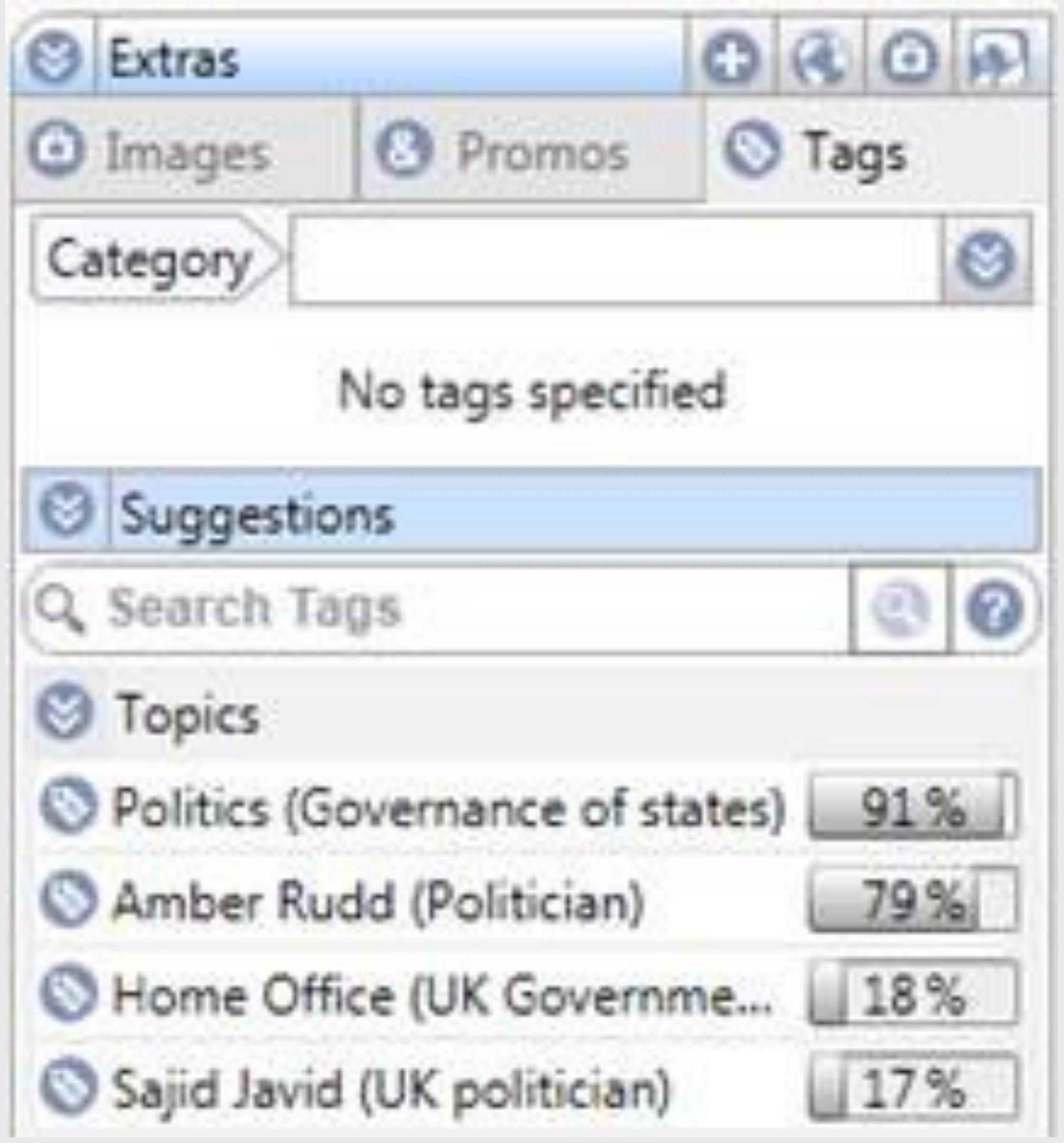
'Wireless' Phones and the House of the Future
DURATION: 03:32

How to Send an Email
DURATION: 01:46

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Tag Suggestion for News and Sport

- BBC journalists manually tag every published article with their primary topics in our Content Production System
- The tags affect where a story might appear
- Automated tag suggestion is used to improve working efficiency and encourage consistent tagging



No tags specified

Topics

Politics (Governance of states)	91 %
Amber Rudd (Politician)	79 %
Home Office (UK Governme...	18 %
Sajid Javid (UK politician)	17 %

Tag suggestion in CPS

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Starfruit Tagging System

- **Starfruit is an automated tag suggestion system for text**
- **Tags use a dictionary of terms called BBC Things**
  - entities
  - themes
  - storylines

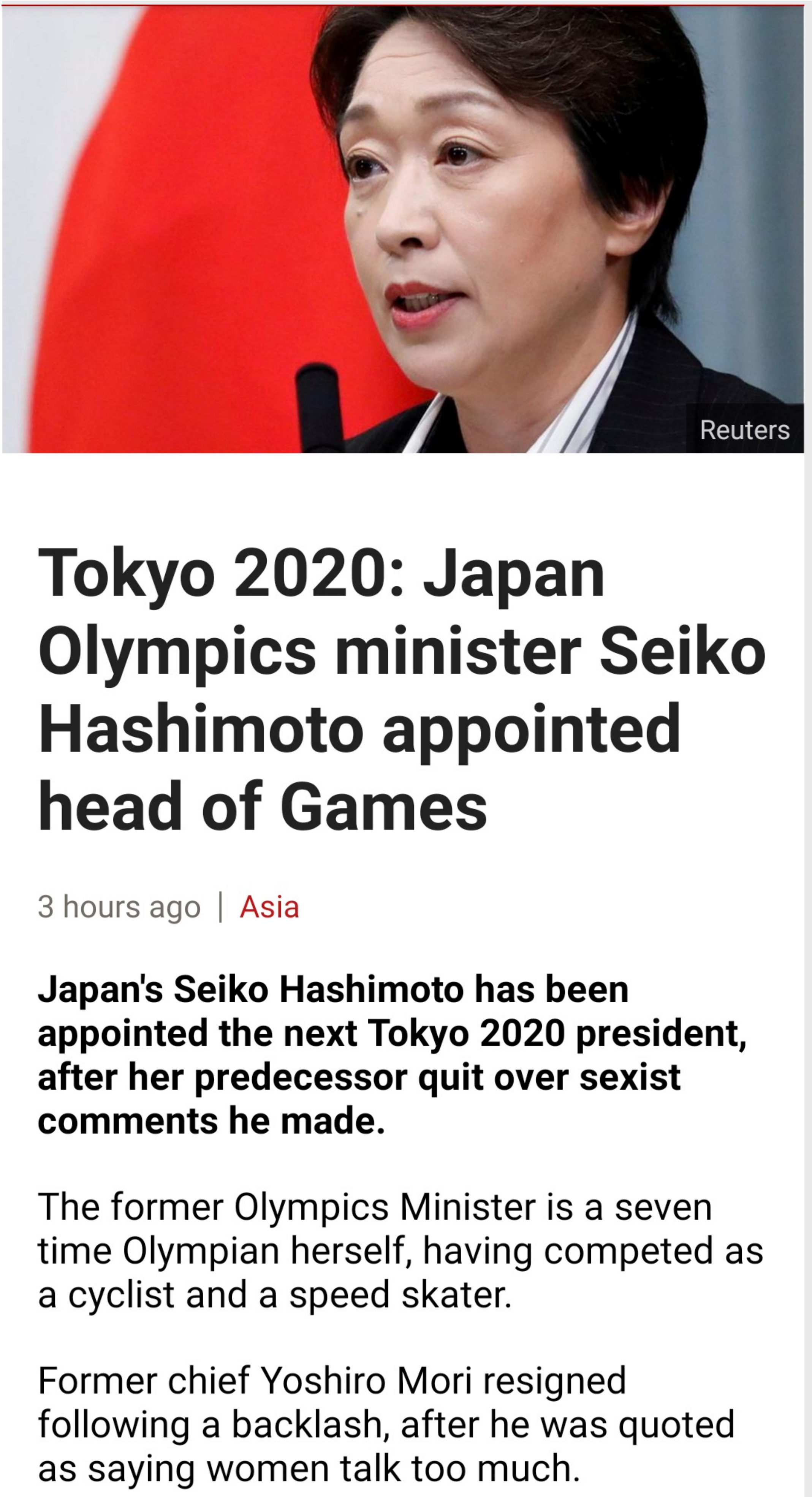
SEMANTIC TAGGING AND TEXT CLASSIFICATION

Starfruit Tagging System

- **Starfruit uses a multilabel text classifier**
  - Trained using 5 years of articles which have been manually tagged by BBC journalists
- **Statistical approach using TF-IDF and Support Vector Machines**
- **11k unique tags**
- **Retrained every 48 hours**

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Starfruit Topics
The primary topics of the text



Reuters

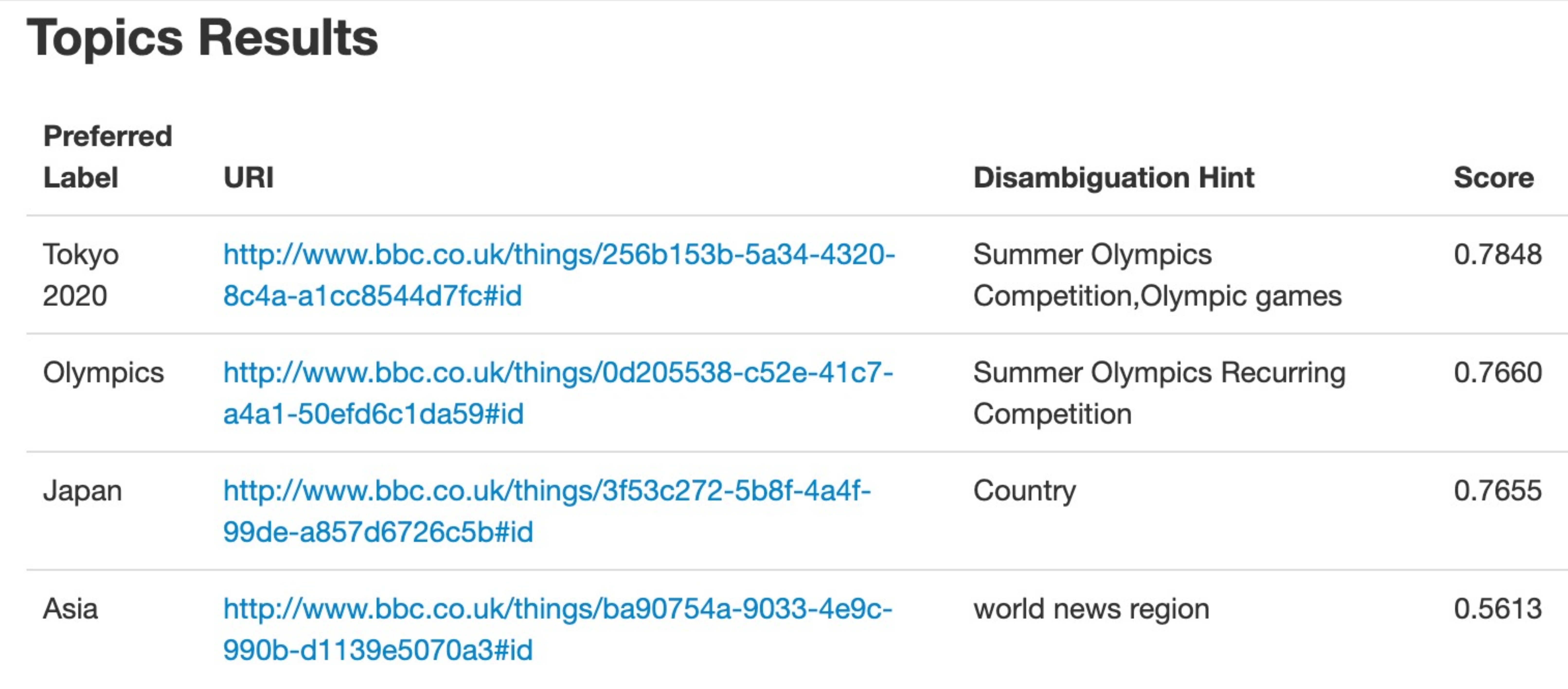
Tokyo 2020: Japan Olympics minister Seiko Hashimoto appointed head of Games

3 hours ago | Asia

Japan's Seiko Hashimoto has been appointed the next Tokyo 2020 president, after her predecessor quit over sexist comments he made.

The former Olympics Minister is a seven time Olympian herself, having competed as a cyclist and a speed skater.

Former chief Yoshiro Mori resigned following a backlash, after he was quoted as saying women talk too much.

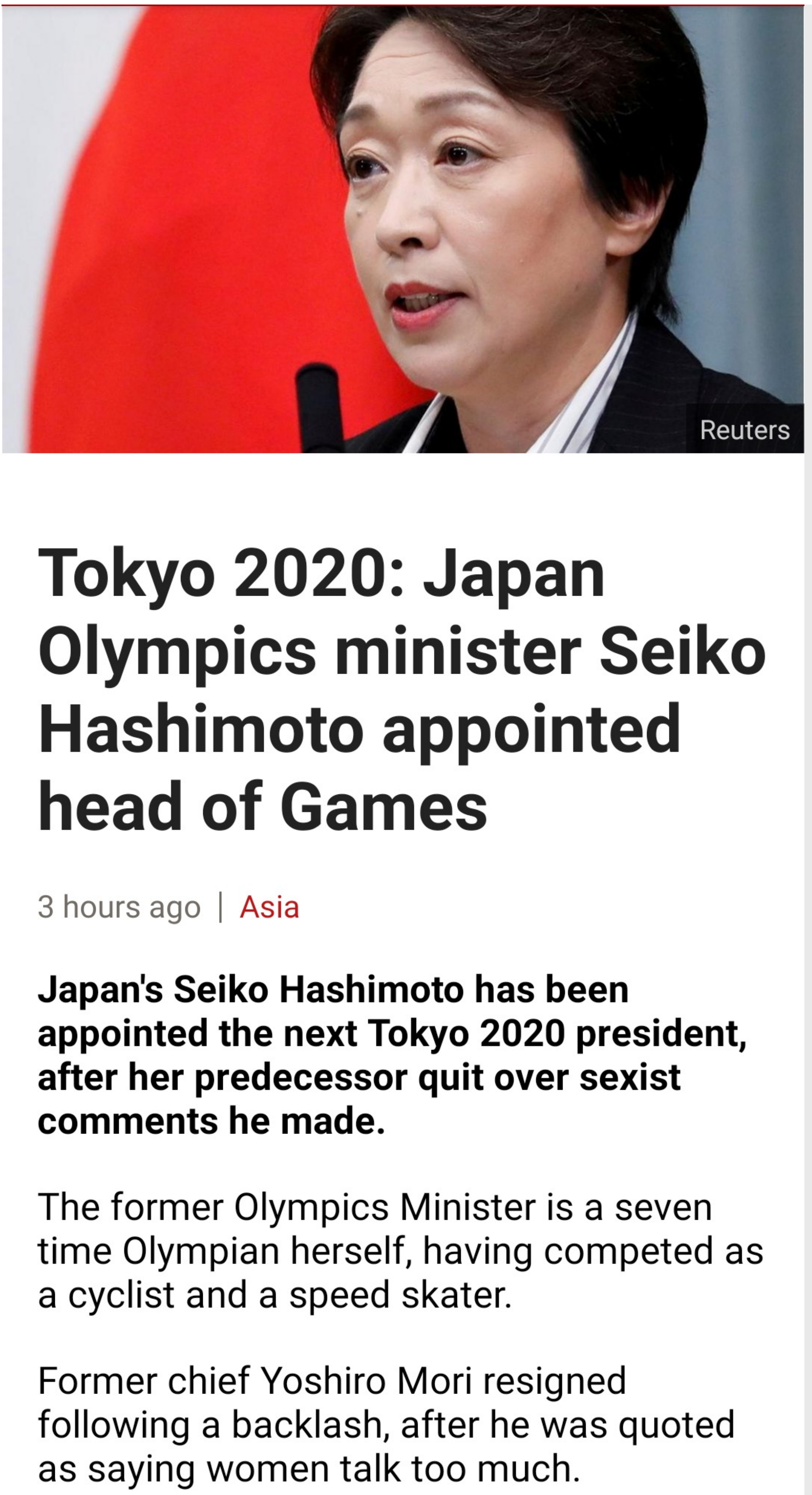


Topics Results

Preferred Label	URI	Disambiguation Hint	Score
Tokyo 2020	http://www.bbc.co.uk/things/256b153b-5a34-4320-8c4a-a1cc8544d7fc#id	Summer Olympics Competition,Olympic games	0.7848
Olympics	http://www.bbc.co.uk/things/0d205538-c52e-41c7-a4a1-50efd6c1da59#id	Summer Olympics Recurring Competition	0.7660
Japan	http://www.bbc.co.uk/things/3f53c272-5b8f-4a4f-99de-a857d6726c5b#id	Country	0.7655
Asia	http://www.bbc.co.uk/things/ba90754a-9033-4e9c-990b-d1139e5070a3#id	world news region	0.5613

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Starfruit Mentions
All entities and themes mentioned in the text



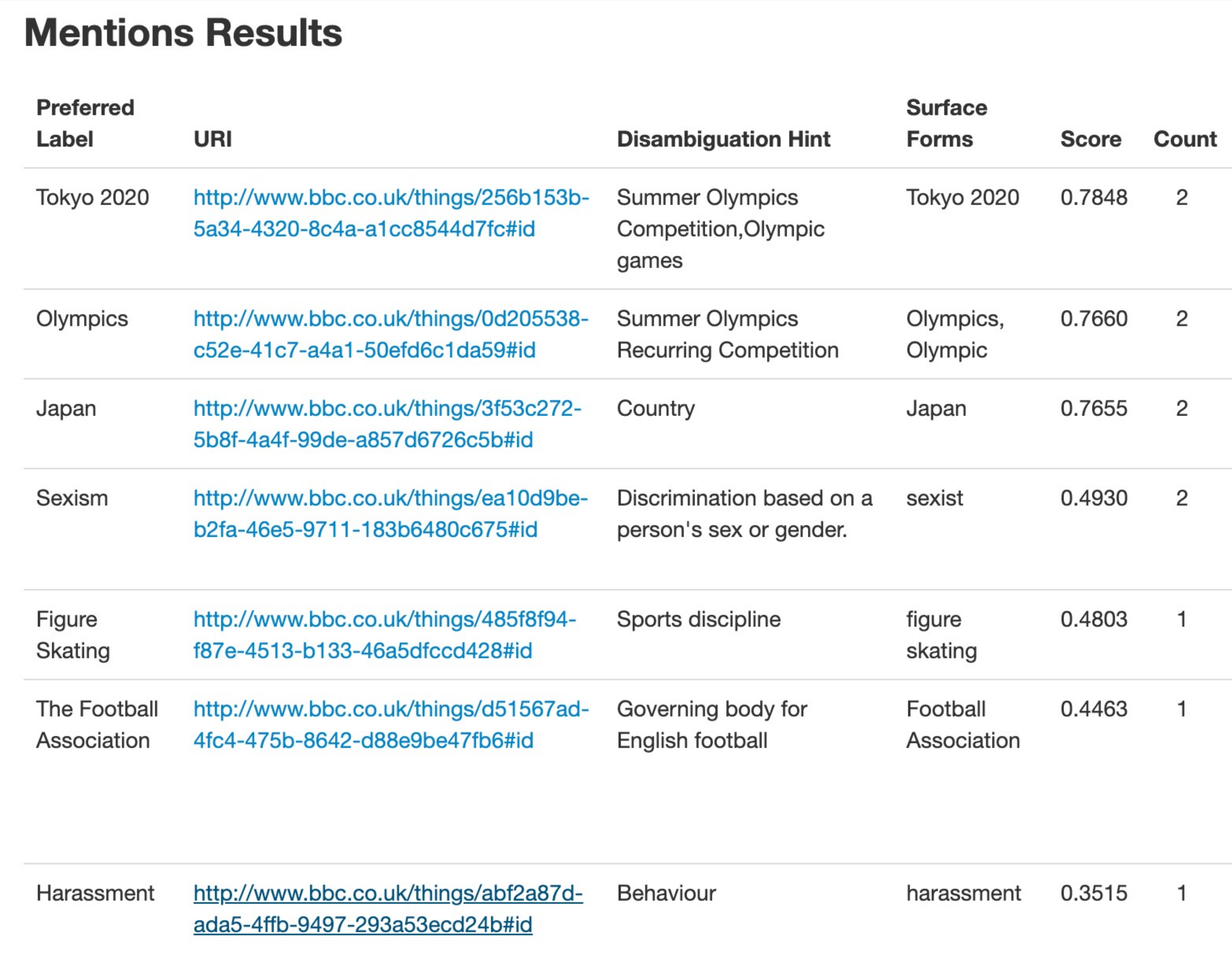
Tokyo 2020: Japan Olympics minister Seiko Hashimoto appointed head of Games

3 hours ago | Asia

Japan's Seiko Hashimoto has been appointed the next Tokyo 2020 president, after her predecessor quit over sexist comments he made.

The former Olympics Minister is a seven time Olympian herself, having competed as a cyclist and a speed skater.

Former chief Yoshiro Mori resigned following a backlash, after he was quoted as saying women talk too much.



Mentions Results

Preferred Label	URI	Disambiguation Hint	Surface Forms	Score	Count
Tokyo 2020	http://www.bbc.co.uk/things/256b153b-5a34-4320-8c4a-a1cc8544d7fc#id	Summer Olympics Competition,Olympic games	Tokyo 2020	0.7848	2
Olympics	http://www.bbc.co.uk/things/0d205538-c52e-41c7-a4a1-50efd6c1da59#id	Summer Olympics Recurring Competition	Olympics, Olympic	0.7660	2
Japan	http://www.bbc.co.uk/things/3f53c272-5b8f-4a4f-99de-a857d6726c5b#id	Country	Japan	0.7655	2
Sexism	http://www.bbc.co.uk/things/ea10d9be-b2fa-46e5-9711-183b6480c675#id	Discrimination based on a person's sex or gender.	sexist	0.4930	2
Figure Skating	http://www.bbc.co.uk/things/485f8f94-f87e-4513-b133-46a5dfccd428#id	Sports discipline	figure skating	0.4803	1
The Football Association	http://www.bbc.co.uk/things/d51567ad-4fc4-475b-8642-d88e9be47fb6#id	Governing body for English football	Football Association	0.4463	1
Harassment	http://www.bbc.co.uk/things/abf2a87d-ada5-4ffb-9497-293a53ecd24b#id	Behaviour	harassment	0.3515	1

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Starfruit Applications

User	System	Application
BBC News	Content Production System	News Articles
Content Distribution	Content Enrichment Chassis	iPlayer Content
BBC News	Wolftech	News Planning

Wolftech Story Tagging Example

https://wolftech.tools.bbc.co.uk
HOME
PLAN
MAP
BOOKING
12:57:19
Wed 17 Feb 2021
Today
0/299 in multiple
Top Stories
Stories
Timeless
Notes
Discarded

STORY

Cancer/Funeral

23 February 2021 | BBC North Ne...

North West | Salim Patel - Journalism 34m ago | Modified by Salim Patel - Journalism 22m ago

Story
Bookings
Publishing points

Description

We did a piece with Sally Cornes from Chorley last autumn. She was "the dancer with cancer" she was ace. Her parents were amazing too. She died last week. She was about to start a clinical trial involving only 50 kids across Europe two days before NYE, but she fell ill on Xmas eve. Her parents wonder if we'd like to interview them & film her funeral today to raise continued importance of research into childhood cancers at a time when funding is plummeting. She had a very active tic-toc account where she posted lots of content. Stuff on Jupiter too.

Additional information

Location/time

+ Add

Tags

Mentions
- **Europe** *world news region*
- **Cancer** *Disease involving unregulated cell gro…*
- **Chorley** *Lancashire, England, United Kingdom*
- **Clinical trials** *Medical and research tests*

About
- **Life Sciences Scotland**
- **Cancer** *Disease involving unregulated cell gro…*
- **Europe** *world news region*

SEMANTIC TAGGING AND TEXT CLASSIFICATION

Alternative Tagging Systems

- DBpedia Spotlight	(open source)
- Tagmatic	(adaptive)
- Amazon Comprehend	
- Google Natural Language	
- Microsoft Azure NLP	
- etc	

(with usual disclaimer)

WOLFTECH STORY TAGGING

Thank you

Chris Newell
chris.newell@bbc.co.uk

Further Information
https://www.bbc.co.uk/rd/projects/natural-language-processing