Contextualizing Trending Entities in News Stories

Marco Ponza, Diego Ceccarelli, Edgar Meij, Paolo Ferragina, Sambhav Kothari

Bloomberg Engineering

The 14th ACM International Conference on Web Search and Data Mining (WSDM 2021)
8-12 March, 2021

Marco Ponza
AI Research Engineer

TechAtBloomberg.com

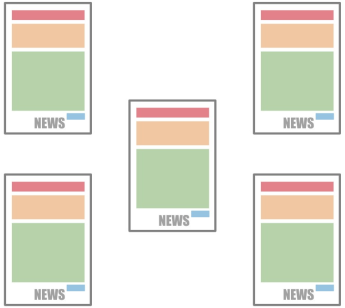
Introduction

TechAtBloomberg.com

Bloomberg

Engineering

Introduction



- Every day an overwhelming amount of data is produced
  - Millions of news stories, web pages, and social media posts

- Important to provide automatic tools for:
  - Identify the most relevant information
  - Understand why a particular piece of information was selected

- Trends distill units (e.g., keywords, phrases, or names) meaningful for characterizing the news stories content
  - Units allow readers to discover and stay focused on relevant information

Introduction

- Bloomberg Terminal functionality:

News Reader Activity | News Sentiment | Twitter Sentiment | News Volume | Twitter Volume | Social Velocity

Largest Increase | Largest Total

	Security	Pub. ↓	GN	Δ Price	Δ AVAT	News Summary
1)	Alibaba Group Holdin...	190		0.00%	+9.83%	Alibaba Group Founder Jack Ma First Appearance
2)	FactSet Research Sys...	112		0.00%	+18.50%	FactSet : 144 01/14/2021
3)	LG Electronics Inc	56		0.00%	+153.32%	17:12 LG Electronics May Exit Smartphone Business
4)	MediaTek Inc	56		+0.23%	+40.94%	Like Qualcomm and Apple, MediaTek Designs Chips
5)	BASF SE	55		+1.24%	+52.13%	BASF Sees FY Adj. Ebit EU3.56 Billion
6)	Tesla Inc	54		0.00%	-34.58%	Tesla Breaking Into India After China Success
7)	Burberry Group PLC	53		+4.43%	+887.53%	Burberry Sales Fall
8)	Taiwan Semiconducto...	53		+0.47%	+67.62%	ASML Beats Grapples With Chip Supply Shortage
9)	Alphabet Inc	50		0.00%	+32.49%	Trump Pardons Former Google Car Engineer
10)	Microsoft Corp	48		0.00%	-3.99%	Microsoft GM Cars

- Can we do one step further and contextualize the trending entities?

Bloomberg Engineering

Introduction

- Can we do one step further and contextualize the trending entities?

  What do we mean for contextualization?

  - Given a trending entity, we want to retrieve and rank contextual entities
  - Contextual entity is an entity that helps explaining why the trending entity is actually trending

- Result of such contextualization can have different applications
  - Entity-centric summary driven by trending and contextual entities
  - Recommend related entities or news stories
  - Query expansion
  - *And many others!*

Our Contributions

New research problem!

Unsupervised and supervised solutions based on:

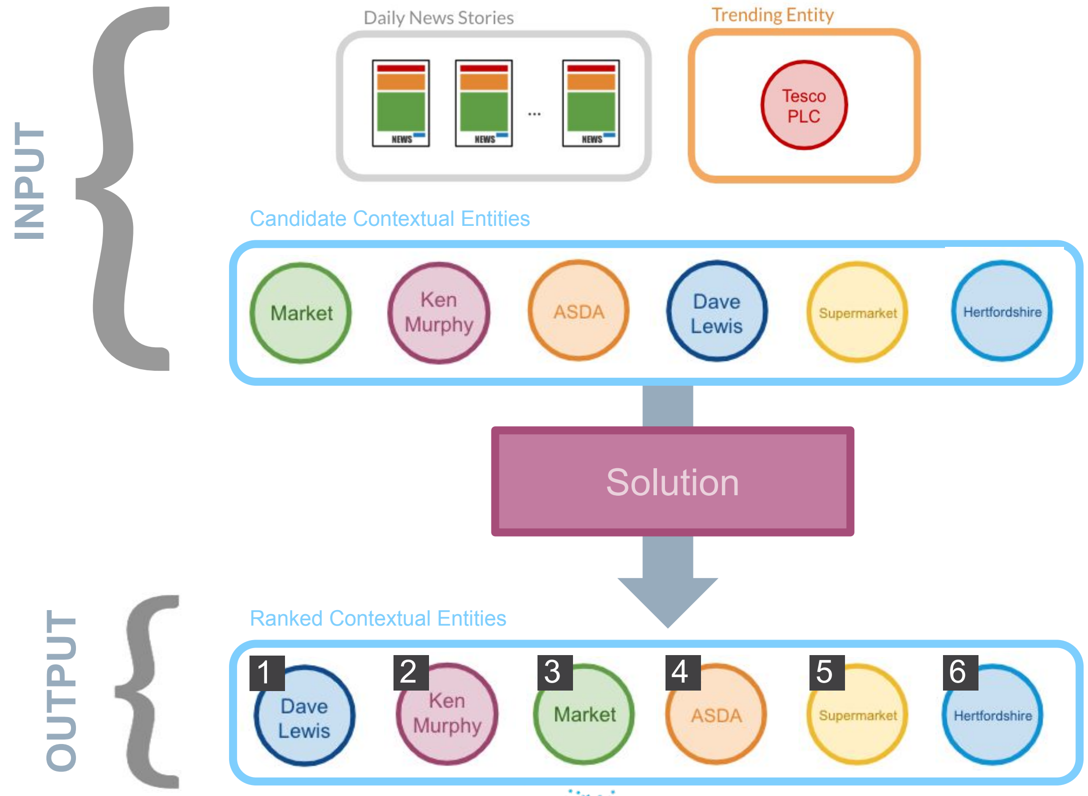
1. Personalized PageRank and entity embeddings
2. Feature engineering and learning to rank

with improvements ranging from 7% to 12% in terms of Precision@1

Creation and release of a test collection built with crowdsourcing. Available at
🔗 https://doi.org/10.5281/zenodo.4422044

Problem Formulation

Contextualization as a Ranking Problem



The CEO of Tesco Dave Lewis resigned and Ken Murphy is supposed to become the new CEO.

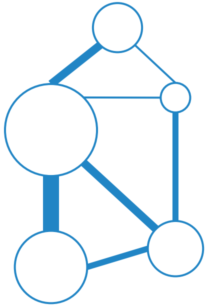
Solutions

TechAtBloomberg.com

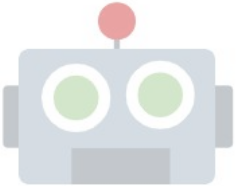
Bloomberg
Engineering

Two Main Solutions

Unsupervised vs. Supervised

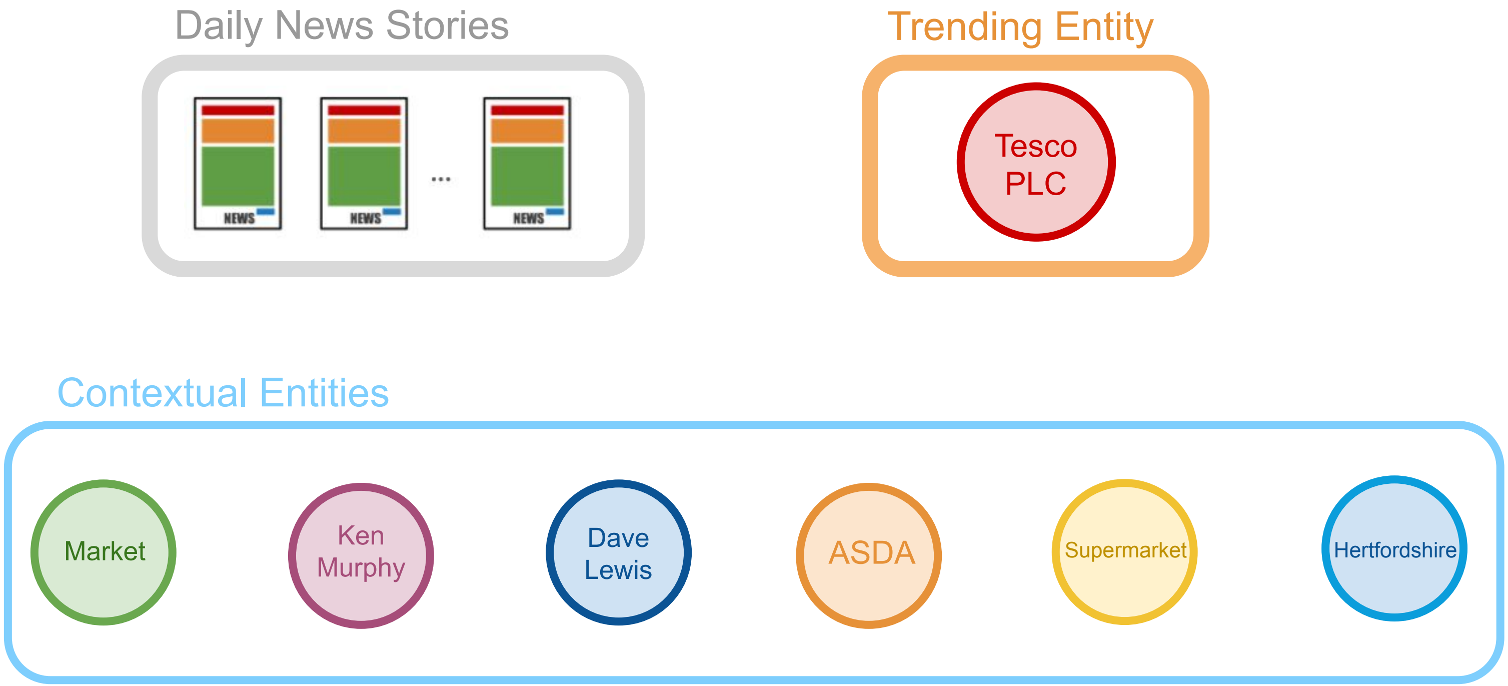


Unsupervised Solution based on Personalized PageRank, Salience, and Embeddings



Supervised Solution based on Feature Engineering and Learn to Rank

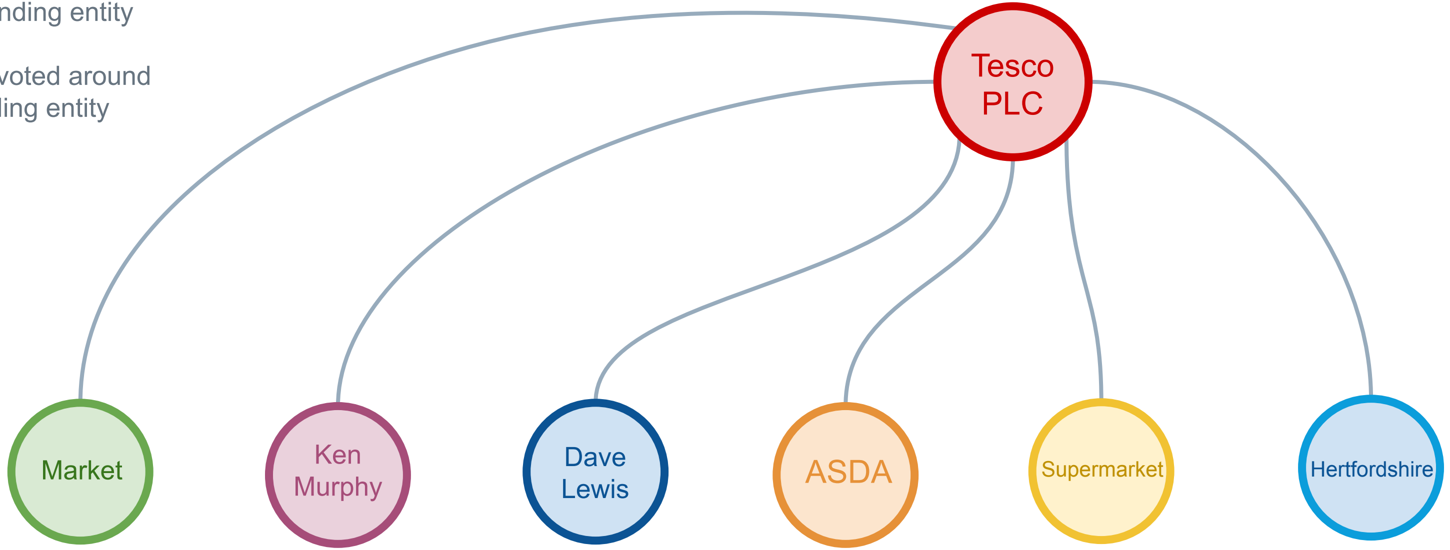
Unsupervised Solution



Unsupervised Solution

Schema: All contextual entities linked to the trending entity

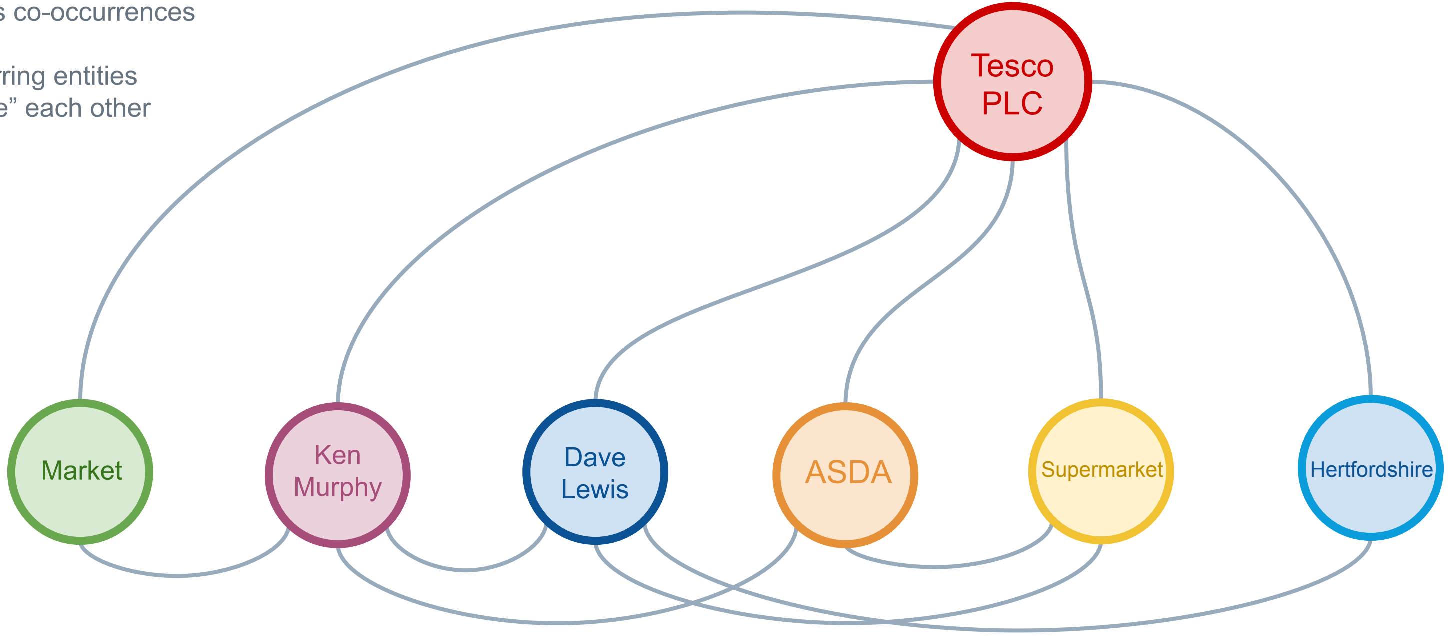
Intuition: Graph pivoted around the trending entity



Unsupervised Solution

Schema: Edges connections draw by stories co-occurrences

Intuition: Co-occurring entities will “vote” each other



TechAtBloomberg.com

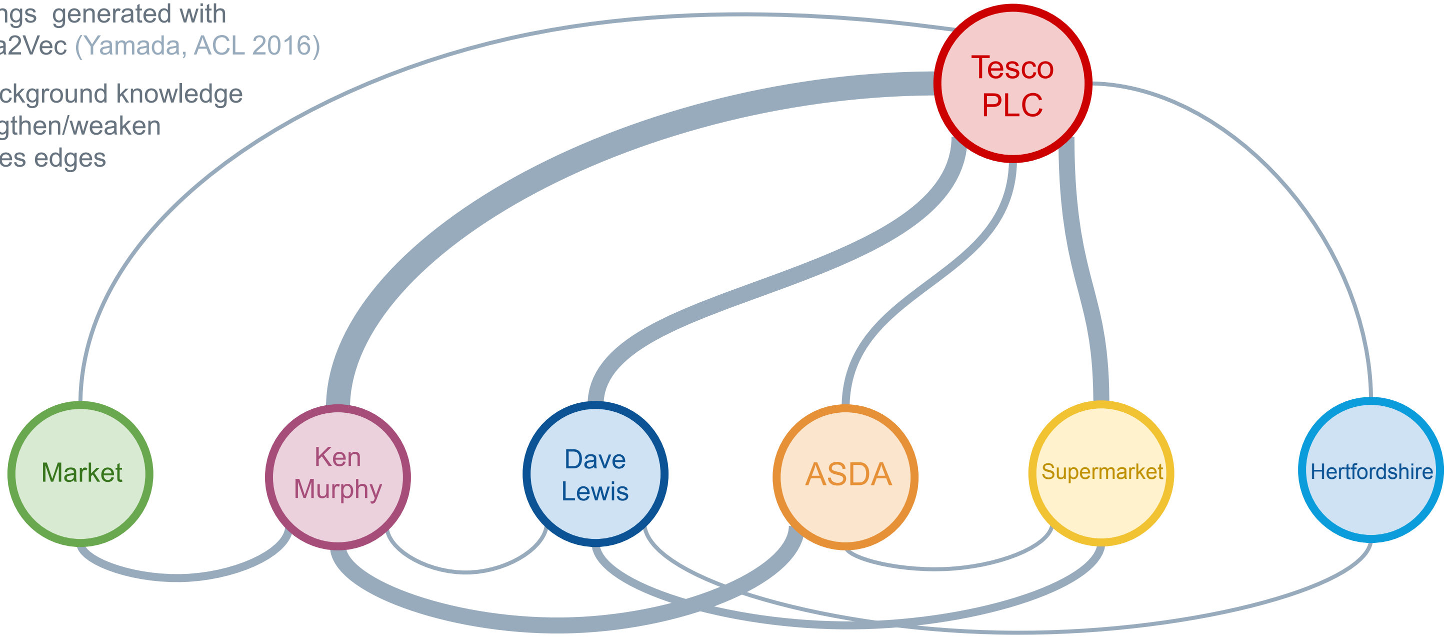
Bloomberg Engineering

Unsupervised Solution

3

Schema: Edges weighted via cosine embeddings generated with Wikipedia2Vec (Yamada, ACL 2016)

Intuition: Using background knowledge for strengthen/weaken the entities edges



TechAtBloomberg.com

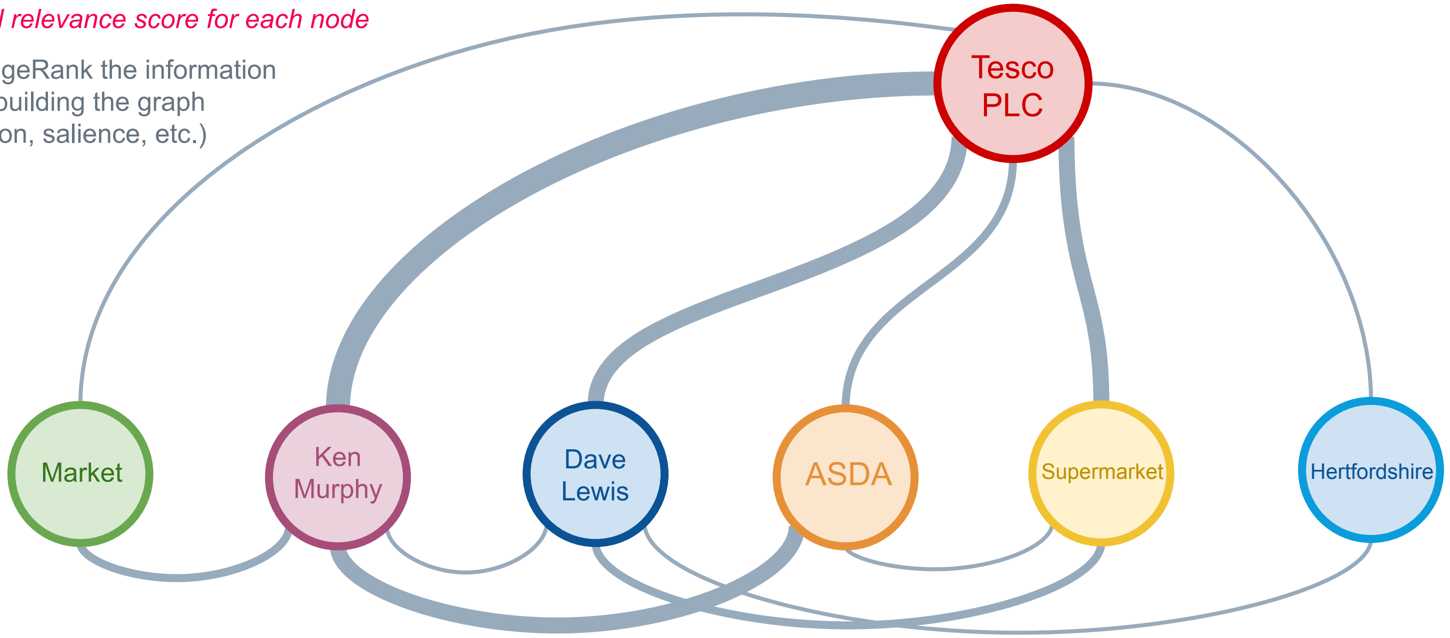
Bloomberg Engineering

Unsupervised Solution

Node Ranking Problem! We can use ***Personalized PageRank***!

Teleport Vector: Initial relevance score for each node

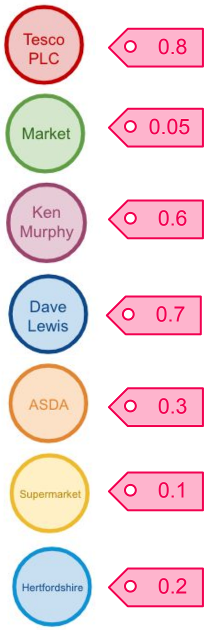
Intuition: Inject in PageRank the information we lost by building the graph (e.g., position, salience, etc.)

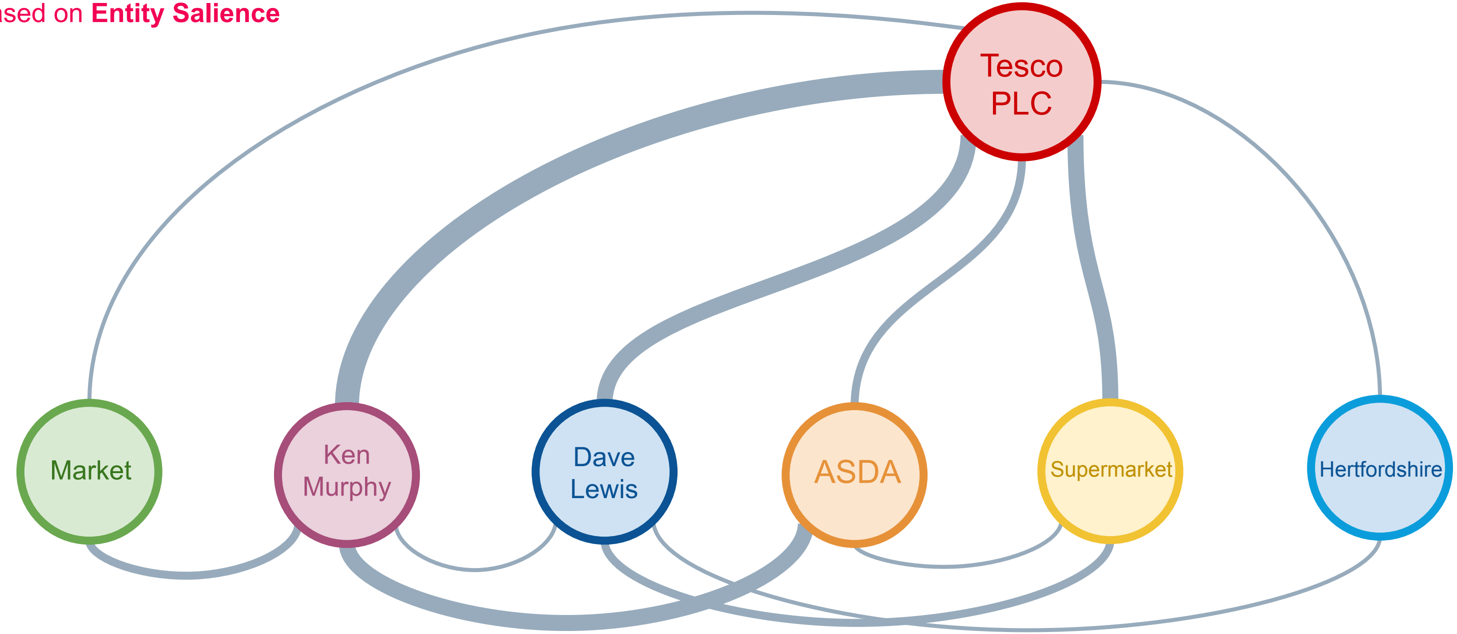


Unsupervised Solution

Node Ranking Problem! We can use ***Personalized PageRank***!

Teleport Vector based on **Entity Salience**





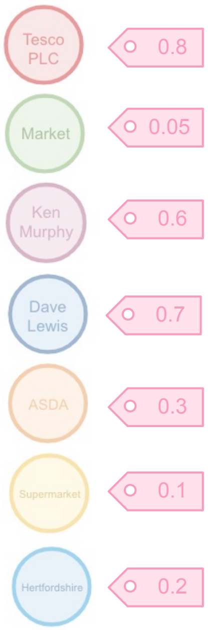
TechAtBloomberg.com

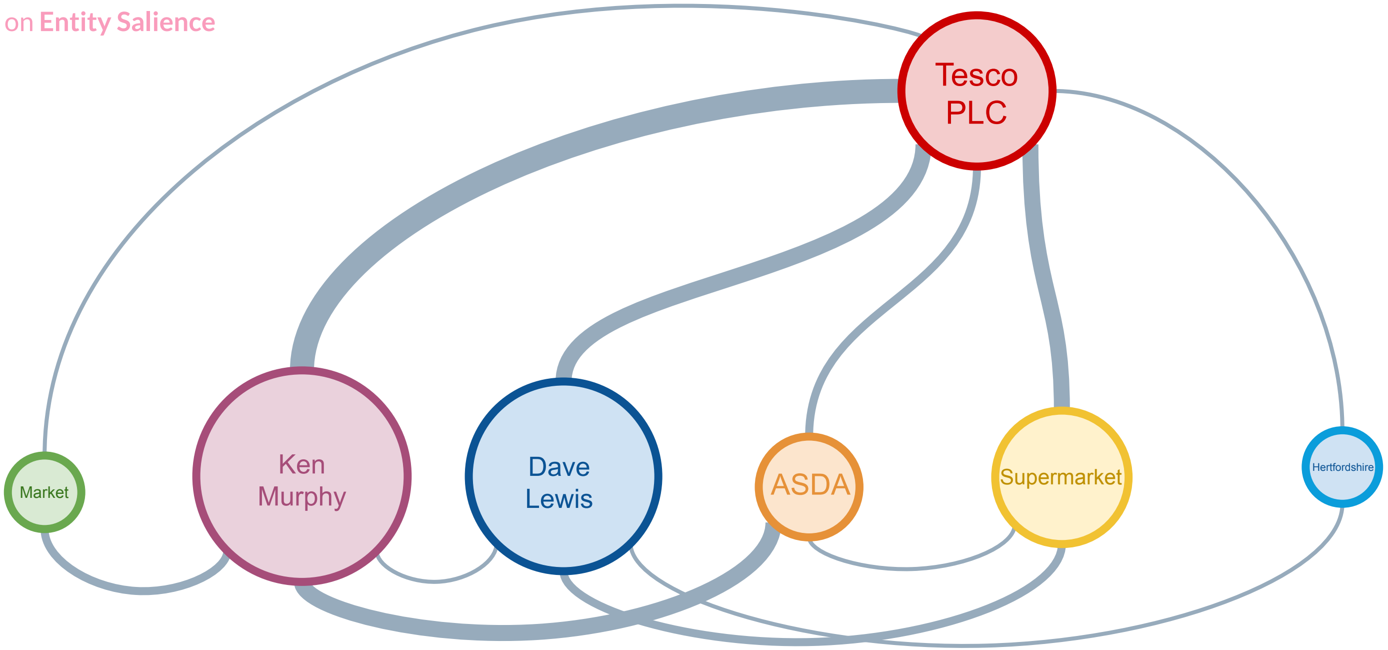
Bloomberg
Engineering

Unsupervised Solution

Node Ranking Problem! We can use ***Personalized PageRank***!

Teleport Vector based on **Entity Salience**





Experimental Results

- Dataset built via crowdsourcing
- 149 trends ~120K contextual entities (~800 per trend)

Publicly available at https://doi.org/10.5281/zenodo.4422044

Experimental Results

	"Relevant" & "Somewhat Relevant" as Gold Labels						"Relevant" as Gold Label					
Method	**MAP**	**P@1**	**P@3**	**NDCG@5**	**NDCG@10**	**MRR**	**MAP**	**P@1**	**P@3**	**NDCG@5**	**NDCG@10**	**MRR**
Frequency	0.098	0.262	0.224	0.168	0.233	0.448	0.097	0.208	0.177	0.179	0.242	0.382
Position	0.237	0.195	0.152	0.237	0.319	0.354	0.247	0.114	0.105	0.249	0.331	0.274
Co-Occurrence	0.359	0.477	0.295	0.441	0.479	0.604	0.441	0.416	0.221	0.486	0.515	0.528
PMI	0.147	0.161	0.15	0.186	0.209	0.324	0.173	0.107	0.105	0.195	0.219	0.252
Milne&Witten	0.177	0.141	0.136	0.179	0.242	0.311	0.177	0.094	0.087	0.174	0.24	0.23
Jaccard	0.214	0.248	0.183	0.234	0.276	0.394	0.229	0.174	0.116	0.240	0.282	0.299
Stories' Embeddings	0.210	0.208	0.161	0.238	0.287	0.373	0.237	0.148	0.110	0.253	0.299	0.295
Wikipedia Embeddings	0.206	0.221	0.154	0.214	0.274	0.372	0.213	0.154	0.096	0.210	0.276	0.276
Reciprocal Rank	0.418	0.523	0.291	0.460	0.508	0.630	0.488	0.430	0.219	0.501	0.542	0.541
Salience	0.497	0.570	**0.394**	0.556	0.612	0.727	0.555	0.456	**0.286**	0.593	0.640	0.622
PPR	**0.519**	**0.644**	0.391	**0.586**	**0.637**	**0.773**$^{\triangle}$	**0.605**$^{\blacktriangle}$	**0.564**$^{\blacktriangle}$	0.282	**0.639**$^{\triangle}$	**0.678**$^{\triangle}$	**0.686**$^{\triangle}$

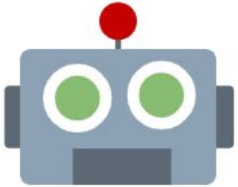
Average improvements around +3/10% wrt Salience

- PPR better for detection of Highly Relevant Entities
- Qualitative better ranking for top ranked entities

Two Main Solutions

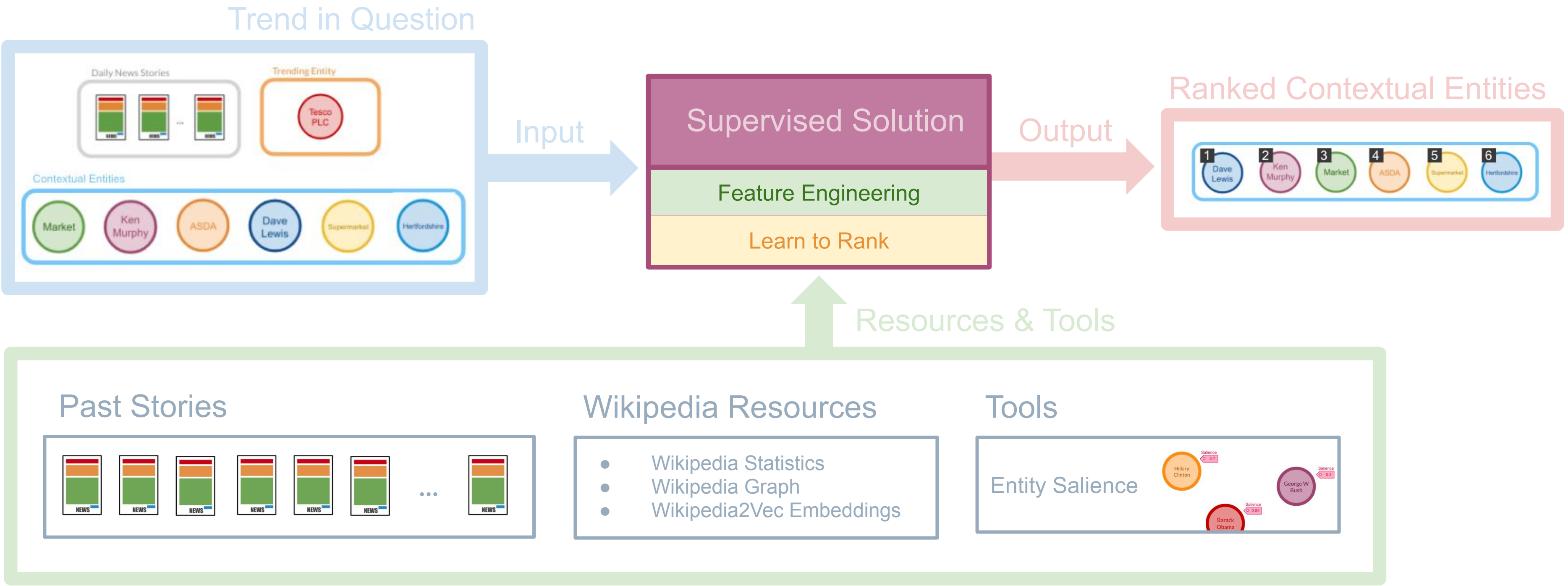
Unsupervised vs. Supervised

Unsupervised Solution based on Personalized PageRank, Salience, and Embeddings



Supervised Solution based on Feature Engineering and Learn to Rank

Supervised Solution

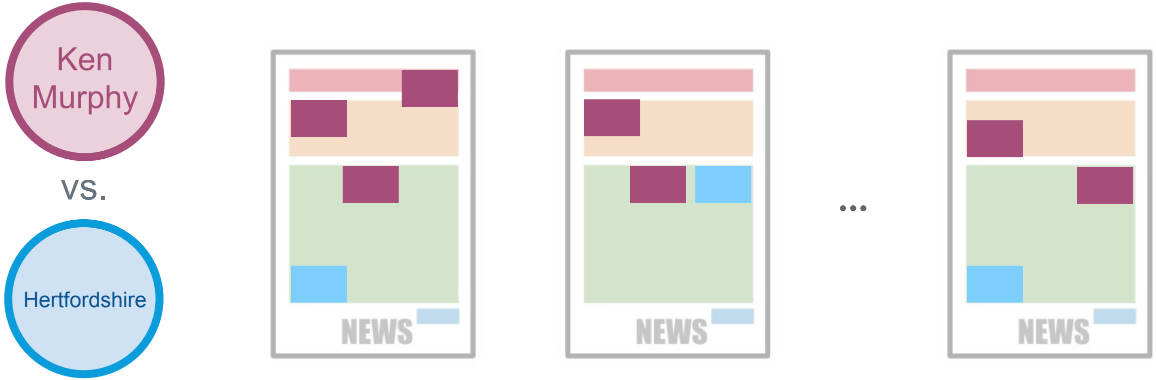


Bloomberg Engineering

Supervised Solution
Feature Engineering

FREQUENCY

*Provide a sense on **how often** an entity is **mentioned** across the news stories*

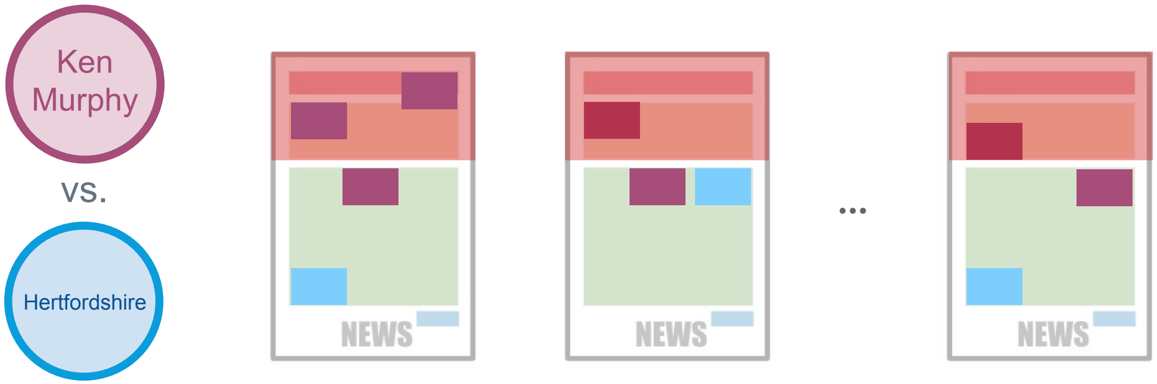


Supervised Solution

Feature Engineering

POSITION

*Provide a sense on **where** an entity is **mentioned** across the news stories*



Supervised Solution

Feature Engineering

POPULARITY

*Provide a sense on **general popularity** of an entity from background knowledge*

Ken Murphy

vs.

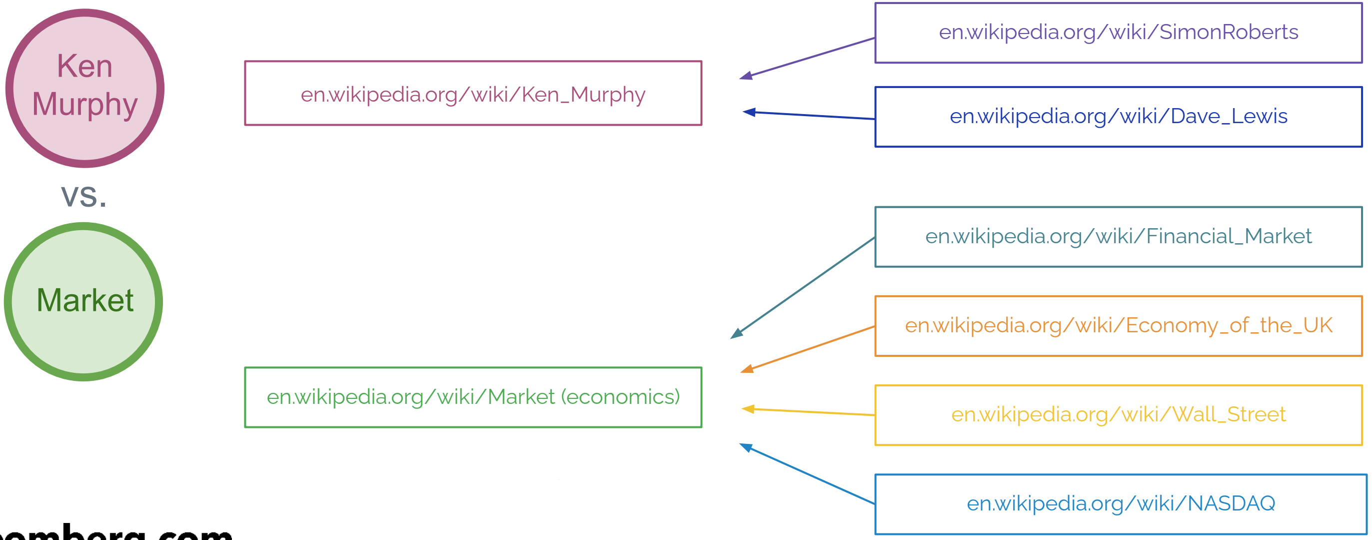
Market

→ Wikipedia

Supervised Solution
Feature Engineering

POPULARITY

Provide a sense on ***general popularity*** *of an entity from background knowledge*



TechAtBloomberg.com

Supervised Solution

Feature Engineering

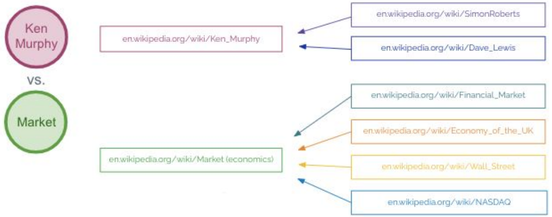
POPULARITY

*Provide a sense on **general popularity** of an entity from background knowledge*

Ken Murphy → *Wikipedia In-Degree* → 12

vs.

Market → *Wikipedia In-Degree* → 648



Supervised Solution

Feature Engineering

TEXT COHERENCE

Provide a sense of ***coherence*** *in the texts of the* ***news stories***

Ken Murphy

Tesco PLC

Dave Lewis

Murphy new Tesco CEO

Will Ken Murphy succeed after Lewis?

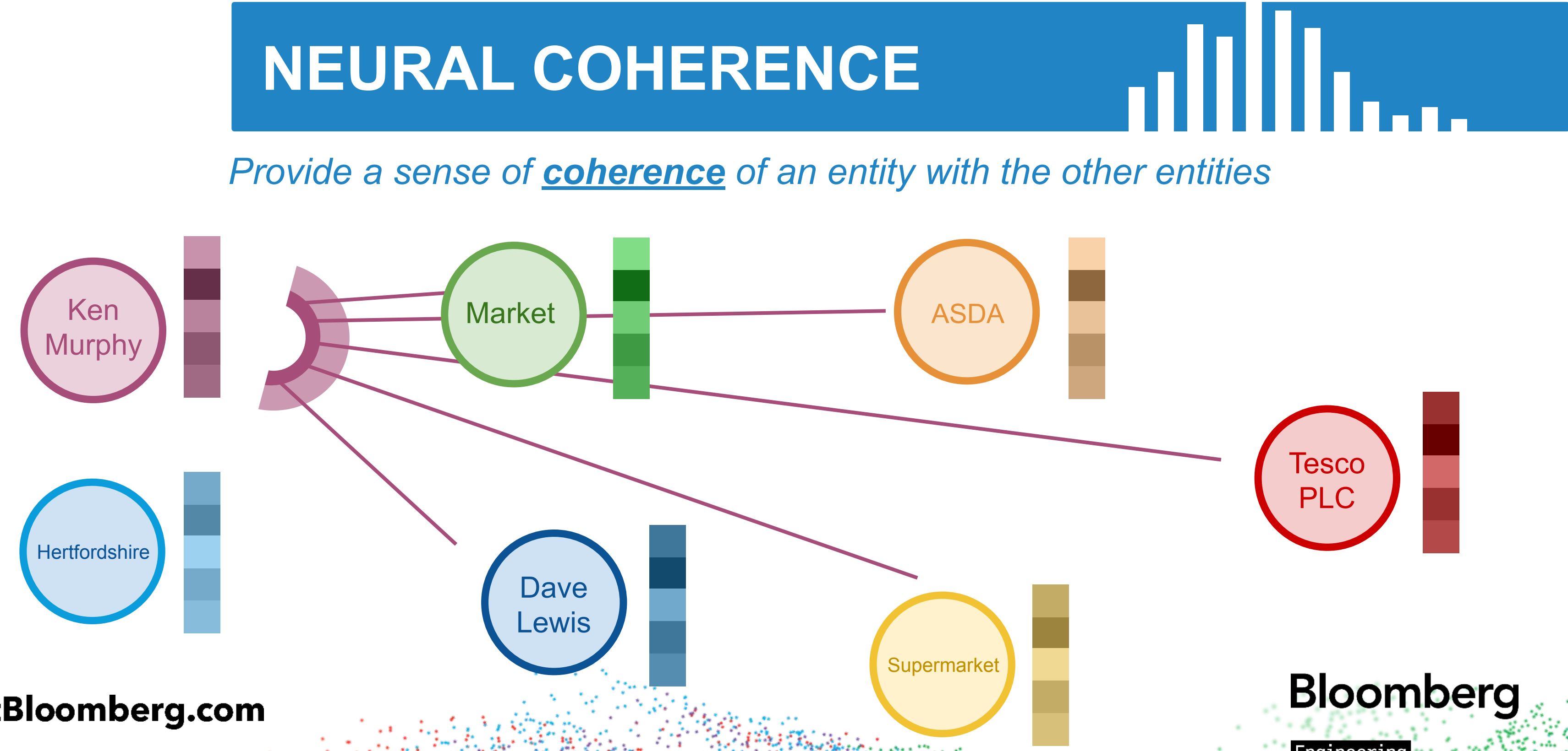
P (Ken Murphy | *Murphy*) = *0.4*

P (Tesco PLC | *Tesco*) = *0.95*

P (Ken Murphy | *Ken Murphy*) = *0.9*

P (Dave Lewis | *Lewis*) = *0.7*

Supervised Solution
Feature Engineering



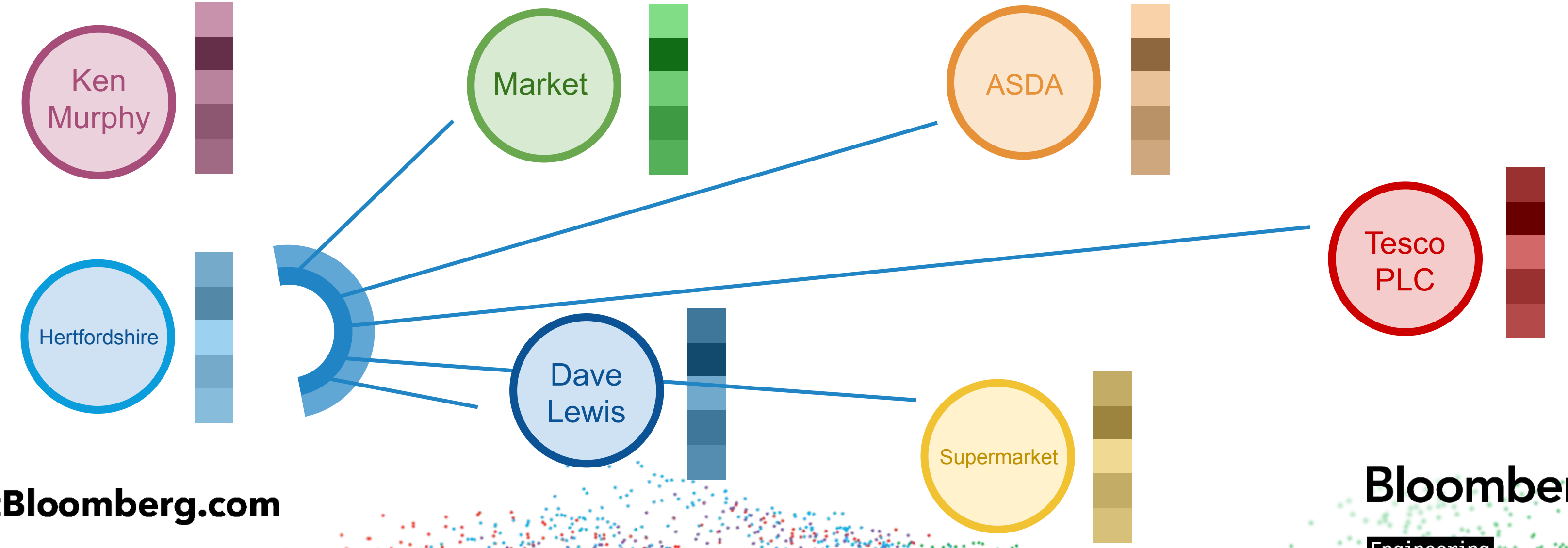
TechAtBloomberg.com

Bloomberg
Engineering

Supervised Solution
Feature Engineering

NEURAL COHERENCE

Provide a sense of __coherence__ of an entity with the other entities



Supervised Solution

Feature Engineering

NEURAL COHERENCE

*Provide a sense of **coherence** of an entity with the other entities*

Ken Murphy → *AVG Embeddings' Cosine* → 0.57

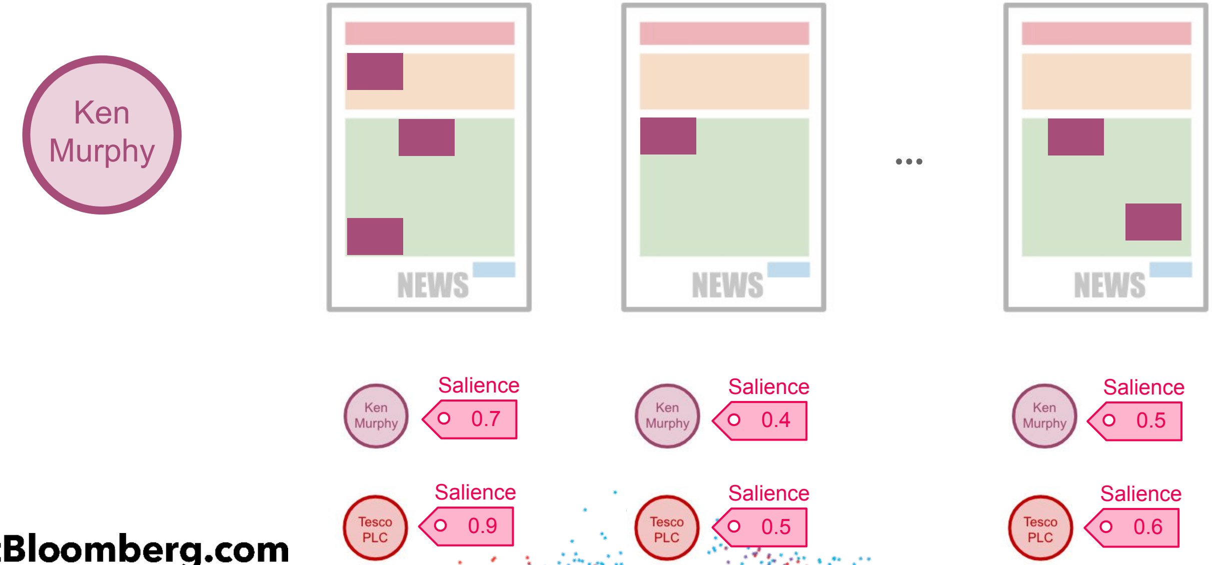
Hertfordshire → *AVG Embeddings' Cosine* → 0.12

Supervised Solution

Feature Engineering

SALIENCE

*Provide a sense on **stories' importance** on an entity*



TechAtBloomberg.com

Bloomberg
Engineering

Supervised Solution

Feature Engineering

SALIENCE

*Provide a sense on **stories' importance** on an entity*

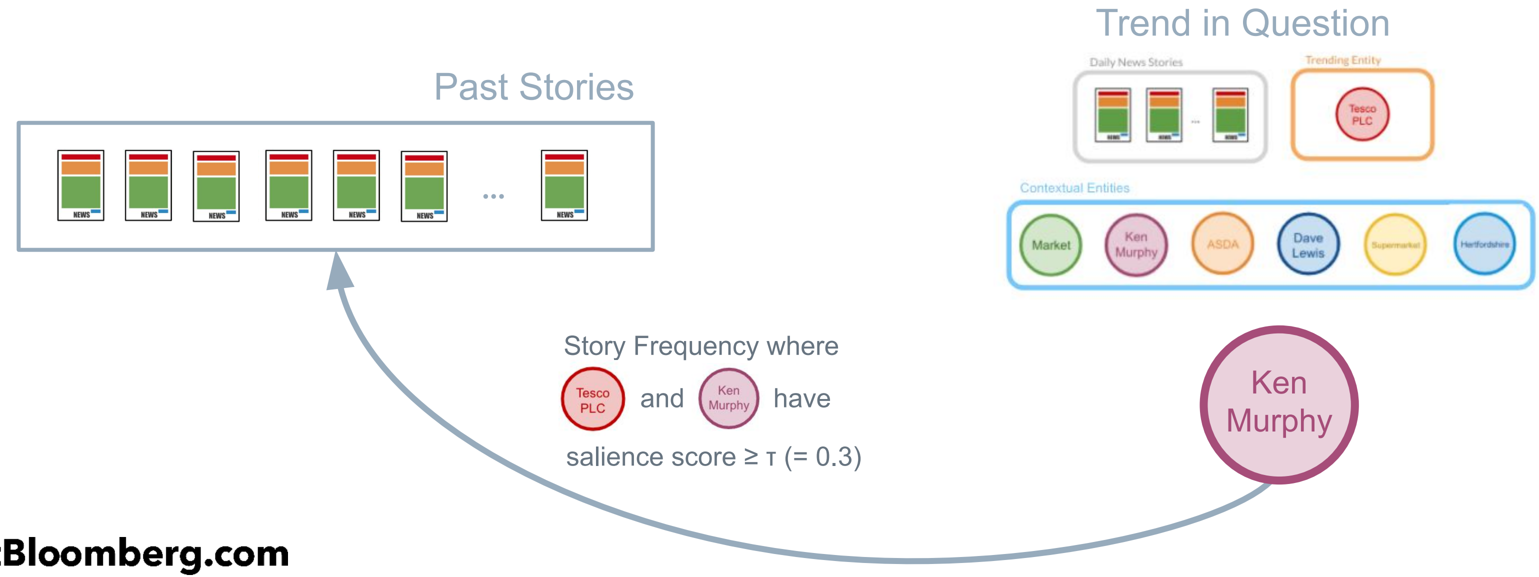
$$\text{Cosine}\left(\begin{array}{|c|c|c|c|}\hline 0.7 & 0.4 & \ldots & 0.5 \\ \hline\end{array},\ \begin{array}{|c|c|c|c|}\hline 0.9 & 0.5 & \ldots & 0.6 \\ \hline\end{array}\right) = 0.73$$

Supervised Solution

Feature Engineering

SALIENCE

Provide a sense on __stories' importance__ on an entity



TechAtBloomberg.com

Supervised Solution

Supervised Solution

Feature Engineering

Learn to Rank

Feature Engineering

Learn to Rank

LightGBM

(Guolin et alii., NeurIPS 2017)

Experimental Results

- Dataset split into training / validation / test sets
- Model tuned on training / validation sets
- Results of the methods reported on the test set

Experimental Results

Method	"Relevant" & "Somewhat Relevant" as Gold Labels						"Relevant" as Gold Label					
	MAP	P@1	P@3	NDCG@5	NDCG@10	MRR	MAP	P@1	P@3	NDCG@5	NDCG@10	MRR
Salience	0.474	0.569	0.364	0.526	0.584	0.714	0.534	0.462	0.251	0.566	0.616	0.604
PPR	0.495	0.646	0.364	0.565	0.617	0.767	0.591	0.554	0.256	0.622	0.659	0.665
LTR	**0.574**◆▲	**0.708**	**0.472**◆▲	**0.629**▲	**0.682**◆▲	**0.815**▲	**0.609**	**0.569**	**0.308**◆▲	**0.654**▲	**0.696**▲	**0.710**▲

Average improvement around +6/9% wrt PPR

- Qualitative better ranking in the head+tail
- Able to detect partially relevant entities

TechAtBloomberg.com

Bloomberg Engineering

Conclusion

TechAtBloomberg.com

Bloomberg

Engineering

Take-Home Messages

- We introduced the problem of ranking contextual entities for a trend

- We experimented with two methods:
  - Unsupervised, based on Personalized PageRank.
  - Supervised, based on feature engineering and learning to rank.

- The Supervised method generally works better than unsupervised ones
  - With improvements between +10% in terms of P@3
  - Feature analysis show most important features are based on salience and text coherence

TechAtBloomberg.com

Bloomberg Engineering

Thank you!

We are hiring: bloomberg.com/engineering

TechAtBloomberg.com