ISE 540
Text Analytics

Mayank Kejriwal
Research Assistant Professor/Research Lead
Department of Industrial and Systems Engineering
Information Sciences Institute
USC Viterbi School of Engineering
kejriwal@isi.edu

MAP vs. NDCG

- Both are important and highly correlated
- NDCG has the advantage that it can take ‘soft’ relevant values into account (MAP can be modified to account for this, but it’s non-standard)
- NDCG and MRR can both be biased in that top-ranked results are given extreme weight
  - To an extent, this problem affects almost all ranking-based IR metrics, by necessity
- **Question:** We had a ‘single-point’ F-Measure between precision and recall to capture the ‘tradeoff’ between the metrics when we used set-based retrieval. Is there a way to capture a similar tradeoff here?

Cont’d

- Just like we computed precision@k, we can also compute recall@k
- Plot a curve!
- Or compute F-Measure@k and then plot it vs. k
- Many ways to get the information you need (remember that it depends on your task)
  - In research, we always use multiple metrics, baselines and benchmarks to make a compelling case

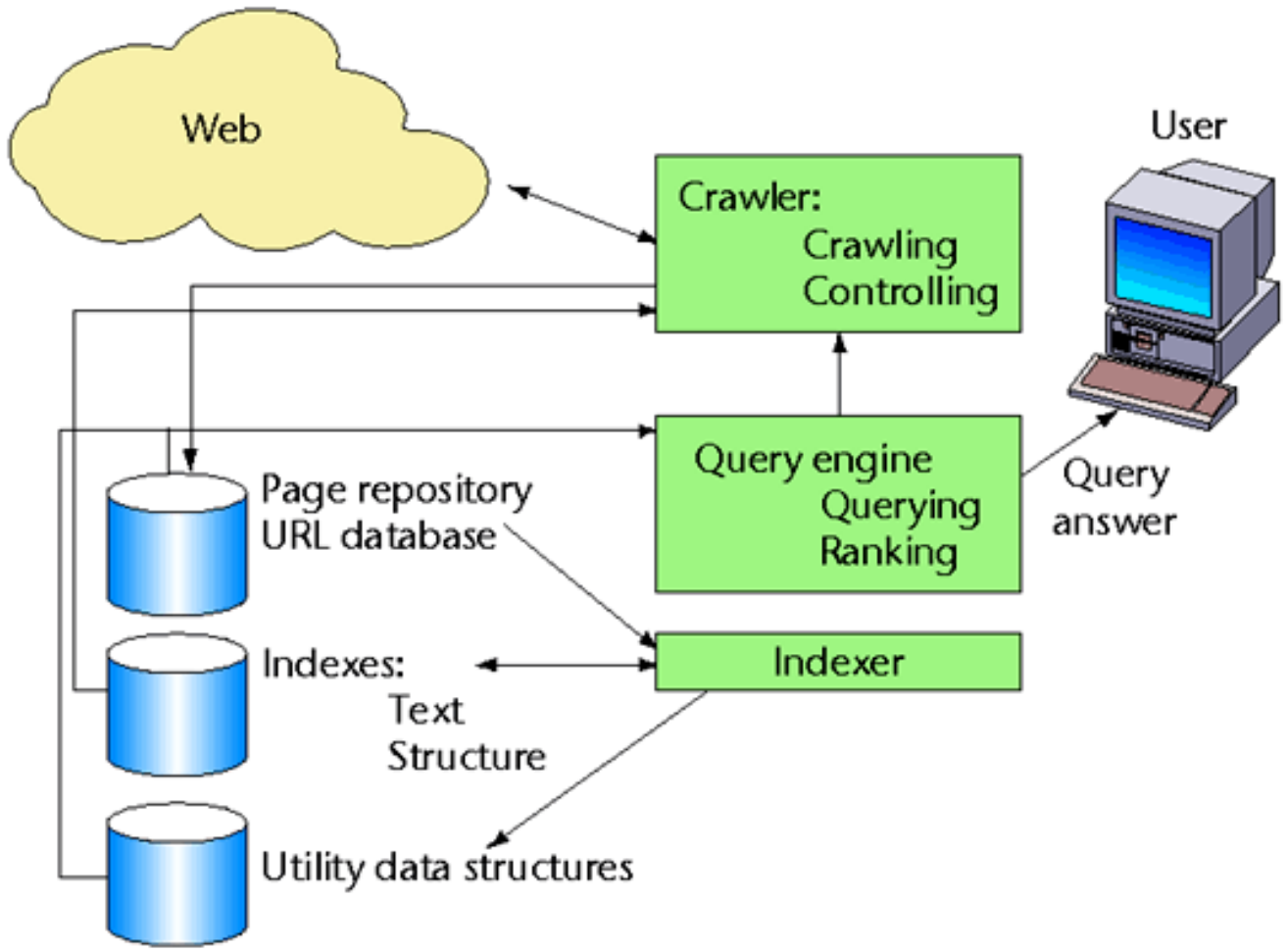
Anatomy of a search engine

Let's try putting all these concepts together

Issues to think about before designing a search engine

- What are the 'queries' and 'documents'?
- For Google, queries are sequences of keywords and documents are webpages that it has managed to crawl and index
- How do we represent queries and documents? Do we represent them in the 'same' vector space (usually yes)?
- As we saw, many options
- Neural models like word2vec, GloVE and fasttext
- Random indexing
- Tf-idf (still the most popular in Information Retrieval, especially if speed and scale are important)
- Given a query and set of documents, all appropriately represented, how do we assign relevance scores to each document?
- How do we evaluate how well we're doing?
- Domain-specific search engines are a hot area of research (e.g., building a search engine over legal documents or social media)
- Could be a project idea

The basics (devil is in the details!)



Word embeddings and representation learning

What's in a word?

— *Three quarks for Muster Mark!*
Sure he hasn't got much of a bark
And sure any he has it's all beside the mark.
But O, Wreneagle Almighty, wouldn't un be a sky o
To see that old buzzard whooping about for uns shi
And he hunting round for uns speckled trousers aro
stown Park?

Quark, one of the most influential of modern Ferengi thanks to his location at Deep Space Nine when the Bajoran wormhole was discovered, owns Quark's Bar on DS9's Promenade, but hates being called a "barkeep," preferring "host" instead as he fancies himself an empathetic dispenser of advice as well as a goodwill ambassador and legitimate entrepreneur extrordinaire.

Quarks and Leptons are the building blocks which build up matter, i.e., they are seen as the "elementary particles". In the present standard model, there are six "flavors" of quarks. They can successfully account for all known mesons and baryons (over 200). The most familiar baryons are the proton and neutron, which are each constructed from up and down quarks. Quarks are observed to occur only in combinations of two quarks (mesons), three quarks (baryons). There was a recent claim of observation of particles with five quarks (pentaquark), but further experimentation has not borne it out.

Quark is similar to French fromage blanc, Indian paneer, and the queso fresco/queijo fresco made in the Iberian Peninsula and in some Latin American countries. It is distinct from Italian ricotta because ricotta (Italian "recooked") is made from scalded whey. Quark is somewhat similar to yogurt cheeses such as the South Asian *chak(k)a*, the Arabic labneh, and the Central Asian suzma or kashk, but while these products are obtained by straining yogurt (milk fermented with thermophile bacteria), quark is made from soured milk fermented with mesophile bacteria.

— *Three quarks for Muster Mark!*
Sure he hasn't got much of a bark
And sure any he has it's all beside the mark.
But O, Wreneagle Almighty, wouldn't un be a sky
To see that old buzzard whooping about for uns sl
And he hunting round for uns speckled trousers a
stown Park?

Quark, one of the most influential of modern Ferengi thanks to his location at Deep Space Nine when the Bajoran wormhole was discovered, owns Quark's Bar on DS9's Promenade, but hates being called a "barkeep," preferring "host" instead as he fancies himself an empathetic dispenser of advice as well as a goodwill ambassador and legitimate entrepreneur extrordinaire.

Quarks and Leptons are the building blocks which build up matter, i.e., they are seen as the "elementary particles". In the present standard model, there are six "flavors" of quarks. They can successfully [illegible] known mesons and baryons (over 200). The most familiar baryons are the proton and neutron, which are each constructed from up and down quarks. Quarks are observed to occur only in combinations of two quarks (mesons), three quarks (baryons). There was a recent claim of observation of particles with five quarks (pentaquark), but further experimentation has not borne it out.

Nothing at all

Quark is similar to French fromage blanc, Indian paneer, and the queso fresco/queijo fresco made in the Iberian Peninsula and in some Latin American countries. It is distinct from Italian ricotta because ricotta (Italian "recooked") is made from scalded whey. Quark is somewhat similar to yogurt cheeses such as the South Asian *chak(k)a*, the Arabic labneh, and the Central Asian suzma or kashk, but while these products are obtained by straining yogurt (milk fermented with thermophile bacteria), quark is made from soured milk fermented with mesophile bacteria.

Firth’s axiom, distributional hypothesis…

“Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache.”

— Ludwig Wittgenstein)

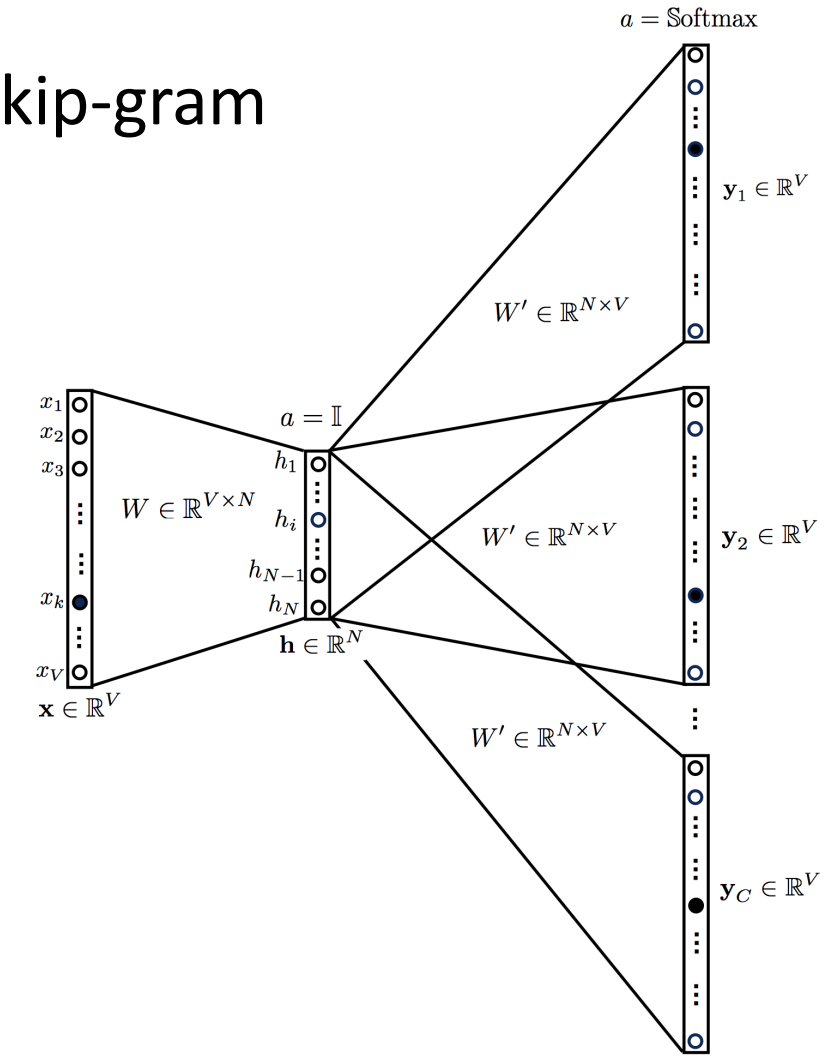
“You shall know a word by the company it keeps!”

— J. R. Firth (1957)

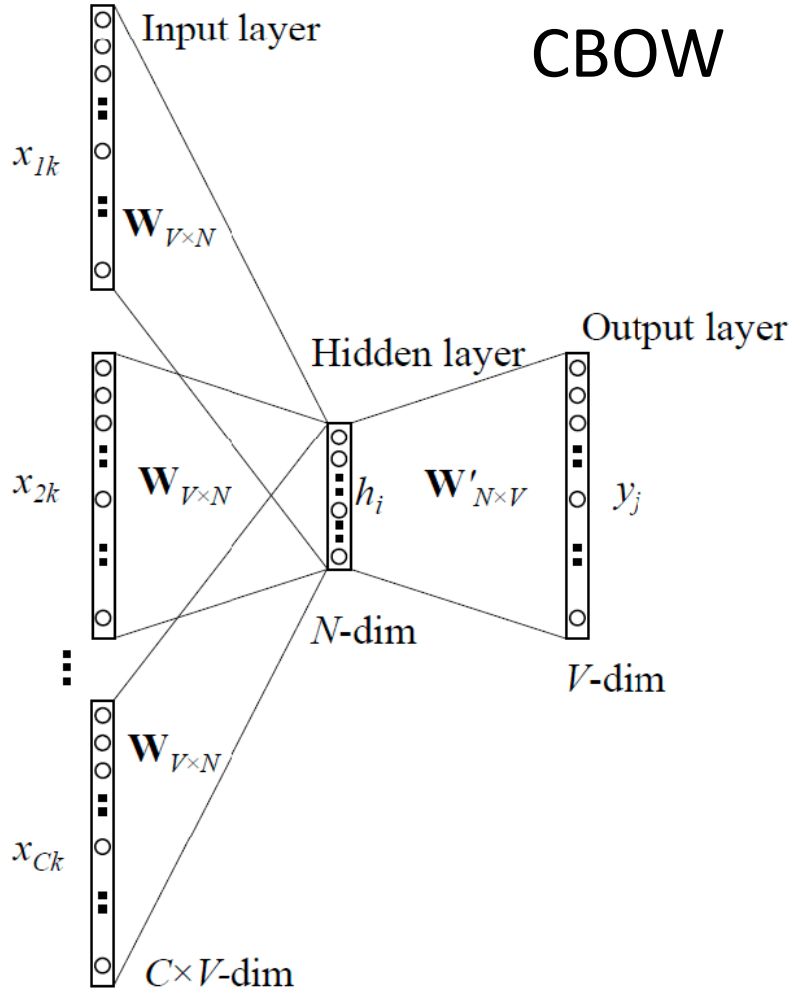
Distributional hypothesis (Zellig Harris 1954)

Inputs and contexts: Skip-gram vs. CBOW vs. other models

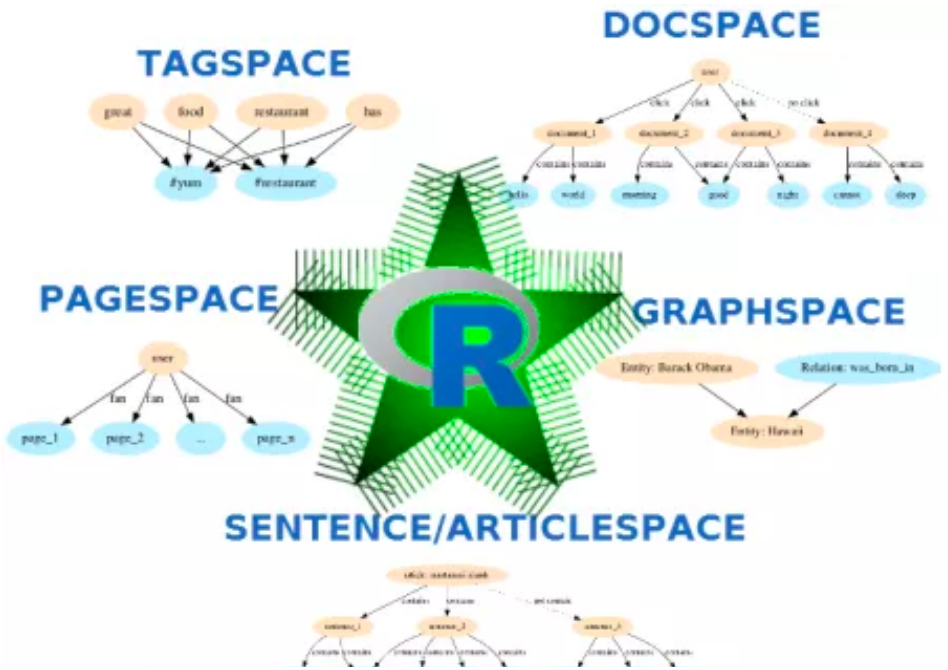
Skip-gram



CBOW



StarSpace



https://github.com/facebookresearch/StarSpace

StarSpace:
Embed All The Things!

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes and Jason Weston
Facebook AI Research

Task	STS12	STS13	STS14	STS15	STS16
fastText (public Wikipedia model)	0.60 / 0.59	0.62 / 0.63	0.63 / 0.62	0.68 / 0.69	0.62 / 0.66
StarSpace [word]	0.53 / 0.54	0.60 / 0.60	0.65 / 0.62	0.68 / 0.67	0.64 / 0.65
StarSpace [sentence]	0.58 / 0.58	0.66 / 0.65	0.70 / 0.67	0.74 / 0.73	0.69 / 0.69
StarSpace [word+sentence]	0.58 / 0.59	0.63 / 0.63	0.68 / 0.65	0.72 / 0.72	0.68 / 0.68
StarSpace [ensemble w+s]	0.58 / 0.59	0.64 / 0.64	0.69 / 0.65	0.73 / 0.72	0.69 / 0.69