

Search Powered by Deep Learning

- From Content Similarity to Semantic Similarity



Speakers



John Kuriakose

Principal Architect
Big Data & Analytics - IIP



**Ashish Vishwas
Kaduskar**

Architect
Big Data & Analytics - IIP



Debanjan Mahata

Senior Research Associate
Big Data & Analytics - IIP

Infosys Information Platform (IIP)

IIP layers

Open Source / Spark Components

ETL / Integration

Spark / Storm / Others

HIVE / HBase/ GraphX / Others

Hadoop / FS Storage / Infra Management

Infosys & Partner IP Components

Tools | Data Extractors | Algorithms | Packaging & Support

Customization, Integration & Implementation Services

Data Modeling & Cleansing | Agile App Development
Data Science & Analytics | Security & Governance
Custom Data Extractors



Agenda

- 
- **What are Word and Doc Embeddings ?**
 - **How do they enrich Search Applications ?**
 - **How to build them for a search application ?**
 - **How to integrate them in a search application ?**

Searching across Research Articles

recurrent neural networks



1. **Recognizing recurrent neural networks (rRNN): Bayesian Inference for recurrent neural networks**

[“Bitzer, Sebastian”, “Kiebel, Stefan J.”] - Fri Jan 20 00:00:00 UTC 2012

[“recurrent neural networks”, “bayesian inference”, “computational neuroscience”, “machine learning applications”, “rnn”, “nonlinear function”, “brain...”

Recurrent neural networks (RNNs) are widely used in computational neuroscience and machine learning applications. In an RNN, each neuron computes its output as a nonlinear function of its integrated input. While the...

2. **Conversion of Artificial Recurrent Neural Networks to Spiking Neural Networks for Low-power Neuromorphic Hardware**

[“Diehl, Peter U.”, “Zarrella, Guido”, “Cassidy, Andrew”, “Pedroni, Bruno U.”, “Nefci, Emre”] - Sat Jan 16 00:00:00 UTC 2016

[“rnn”, “artificial recurrent neural networks”, “low-power neuromorphic hardware”, “neuromorphic low-power systems”, “significant momentum”, “recurrent...”

In recent years the field of neuromorphic low-power systems that consume orders of magnitude less power gained significant momentum. However, their wider use is still hindered by the lack of algorithms that can harness the...

3. **Sequence Modeling using Gated Recurrent Neural Networks**

Pezeshki, Mohammad - Thu Jan 01 00:00:00 UTC 2015

[“gated recurrent neural networks”, “recurrent neural networks”, “human motion data”, “next immediate data point”, “recently proposed gated recurrent units”, “promisi...”

In this paper, we have used Recurrent Neural Networks to capture and model

Ingestion



- Data Preprocessing
- Indexing

Enrichment



- Training different word embedding models

Content Selection

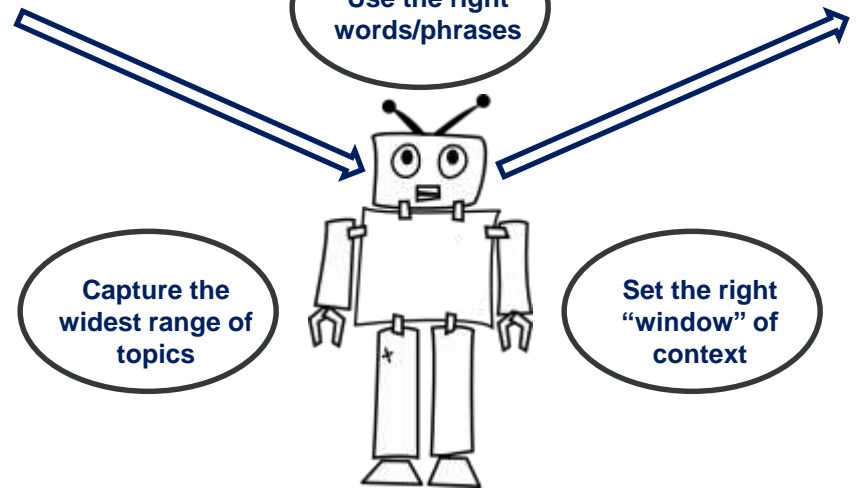


- Query Expansion
- Similar Article Recommendation
- Keyword Extraction

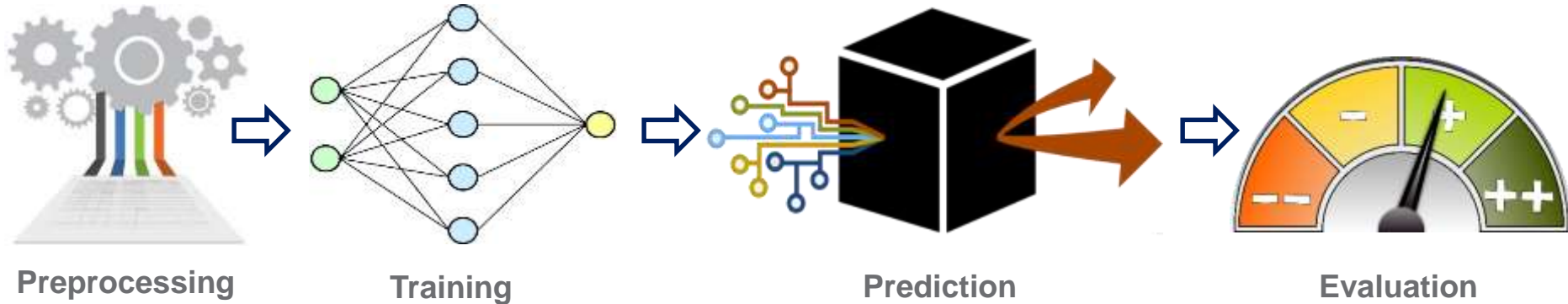
Humans Vs Machines



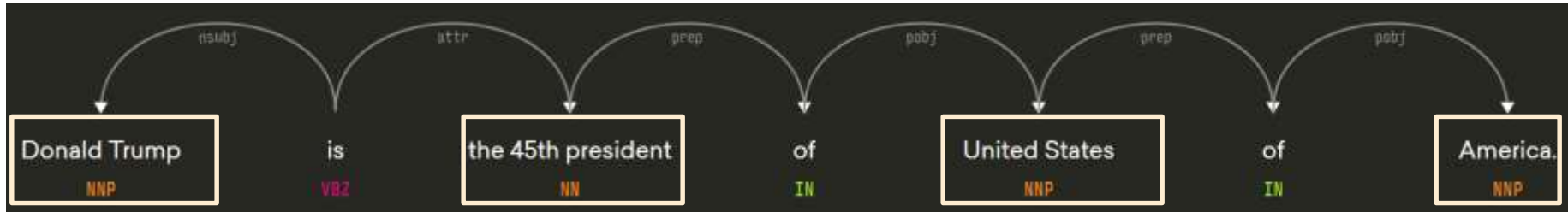
1147001 Research Abstracts



Building the Models



Preprocessing



Donald Trump **PERSON** is the 45th president of United States of America **GPE** .

Donald Trump is the 45th president of United States of America.

Donald Trump 45th president United States of America.

Sentence splitting

Phrase Tokenization

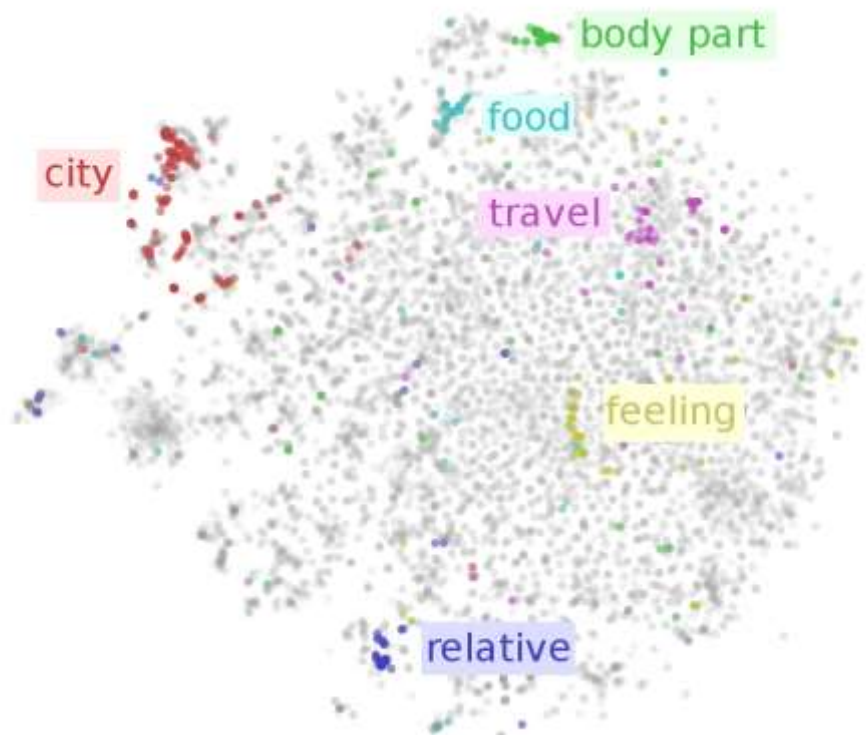
Removal of strings containing only numeric characters

Removal of functional words like 'accordingly', 'although', etc

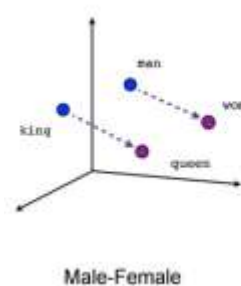
Removal of named entities of types "DATE", "TIME", "PERCENT", "MONEY", "QUANTITY", "ORDINAL", "CARDINAL"



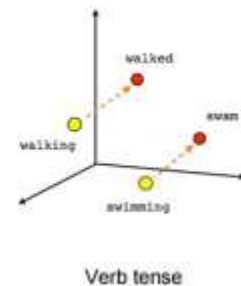
Word Embeddings



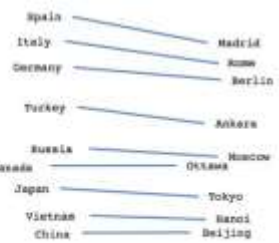
```
array([-2.93477079e-02,  4.84889327e-02,  1.38789033e-02,  1.70784773e-01,  -8.15586085e-02,  8.64981543e-02,  1.66968349e-02,  -4.40942113e-02,
        1.11839613e-01,  1.43412849e-02,  3.57013787e-02,  2.48175624e-02,  2.73818902e-01,  2.38924382e-01,  2.49524382e-01,  2.49524382e-01,
        -3.21225047e-02,  1.68366344e-01,  7.51848706e-02,  9.73189352e-02,  2.91825386e-02,  -8.39889130e-02,  -1.25252504e-02,  -8.39889130e-02,
        -3.06699536e-01,  1.76978416e-02,  2.89128951e-02,  3.45082753e-02,  7.99534091e-02,  -7.14821371e-03,  5.42953612e-02,  -7.89423361e-03,
        1.73390688e-02,  1.83222959e-01,  3.73032984e-02,  3.70388124e-02,  -1.06728123e-02,  1.16309080e-02,  8.88933724e-03,  3.35439409e-02,
        -1.26840089e-01,  2.78647892e-02,  8.88933724e-02,  3.27521139e-02,  2.26881046e-02,  -1.84341872e-01,  -1.84982379e-02,  1.22808434e-01,
        -8.97222534e-02,  -2.69968430e-02,  1.21217552e-02,  1.40727863e-01,  5.01757876e-02,  3.01912884e-02,  4.87424132e-02,  4.87424132e-02,
        7.84821086e-02,  2.48976751e-02,  3.12786886e-02,  -7.53788016e-02,  -9.99753826e-02,  -5.08708926e-02,  -4.43738926e-02,  -5.35182714e-02,
        4.12802517e-02,  3.02307527e-02,  -6.09082734e-02,  9.17636827e-02,  -1.70259083e-02,  1.12309090e-02,  1.10070890e-03,  4.83026084e-02,
        -1.46517027e-02,  8.62820297e-02,  4.43889266e-02,  1.10812598e-02,  7.05480827e-02,  4.73822841e-02,  3.78829528e-02,  8.88888188e-02,
        -6.40487292e-02,  2.24820722e-02,  3.42784189e-02,  -2.38551638e-02,  3.07184184e-03,  1.68418756e-02,  -3.07184184e-02,  -3.27486466e-02,
        1.26415187e-02,  3.16781773e-02,  5.72336986e-02,  6.87443405e-02,  3.08820328e-02,  -8.31050586e-02,  3.07016436e-02,  1.11978104e-01,
        1.18881555e-01,  1.42577818e-01,  5.88884186e-02,  5.54820548e-02,  5.54820548e-02,  -3.28798456e-03,  6.88794556e-03,  -3.48612195e-03,
        1.69888453e-01,  1.35308862e-01,  2.58461349e-02,  2.23174957e-02,  -5.48677238e-02,  -4.88176173e-02,  8.17817399e-02,  1.10389189e-01,
        4.88884711e-01,  1.71381848e-02,  1.57616164e-02,  8.88881088e-02,  4.68536117e-02,  -4.78882444e-02,  5.45848484e-02,  -4.78882444e-02,
        -1.17841928e-02,  -1.47863442e-02,  9.44363581e-02,  2.57888342e-02,  -6.98888173e-02,  -8.44881556e-02,  8.44881556e-02,  8.44881556e-02,
        3.88879983e-02,  -4.87979386e-02,  3.88133231e-02,  1.87884955e-02,  7.18847933e-02,  8.28471752e-02,  7.83251956e-02,  -4.88814956e-02,
        -4.55578883e-02,  2.64884718e-02,  8.88415725e-02,  8.38839968e-02,  1.18888188e-02,  8.42872825e-02,  2.13884778e-02,  9.12971215e-02,
        3.16888448e-01,  -6.42881838e-02,  4.18844711e-02,  9.73825488e-02,  1.38888388e-02,  8.23825388e-02,  -7.13825388e-02,  3.18884563e-01,
        -2.27844847e-02,  1.88813836e-01,  -3.83286777e-02,  -6.83754646e-02,  3.88888688e-02,  3.83252996e-02,  1.81176299e-02,  -1.84858819e-01,
        -6.68888487e-02,  3.78787854e-02,  -4.28888888e-02,  -4.38728237e-02,  8.84281082e-02,  -1.13224725e-02,  9.81313898e-02,  9.81313898e-02,
        1.38888888e-02,  6.78131534e-02,  -2.88718531e-02,  6.83873587e-02,  8.83873587e-02,  -2.13852897e-03,  8.81888246e-02,  -4.28884888e-02,
        -1.83214814e-01,  8.53878888e-02,  7.53884838e-02,  -4.81884838e-02,  9.53884838e-02,  2.13884838e-02,  3.78884838e-02,  4.88884838e-02,
        -7.48884838e-02,  7.01783825e-02,  2.38712382e-02,  -2.88712382e-02,  -5.88712382e-02,  -1.88712382e-02,  1.18884838e-01,  1.88884838e-01,
        2.18788881e-02,  -1.45813554e-02,  -8.28135442e-02,  -6.82858593e-02,  -2.84820838e-02,  2.24213188e-02,  4.87499543e-04,  -7.58888888e-02])
dtype=float32)
```



Male-Female



Verb tense

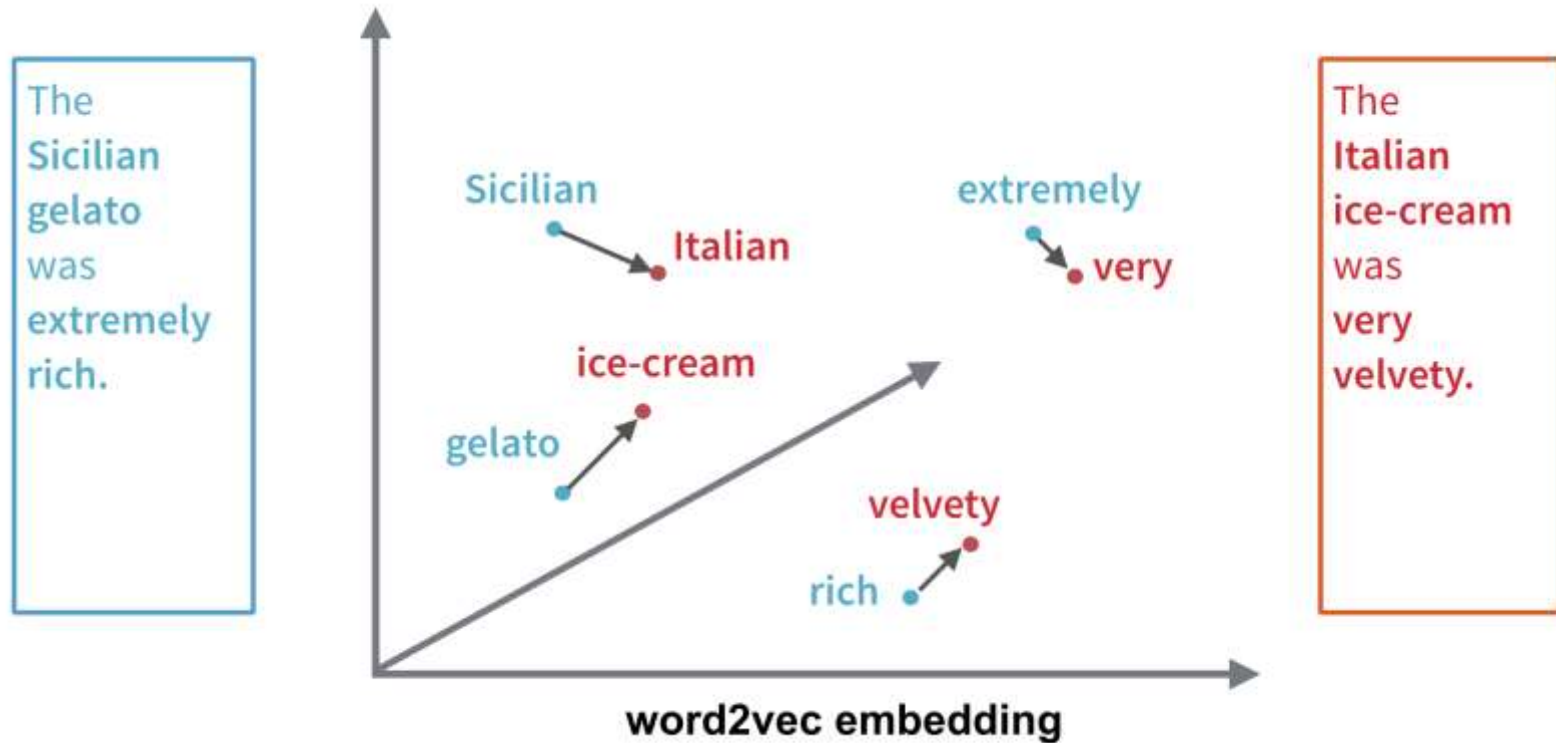


Country-Capital

Source: <http://sebastianruder.com/word-embeddings-1/index.html>

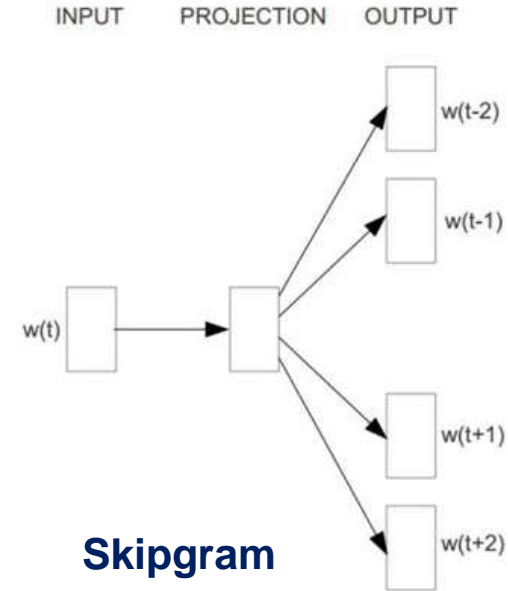
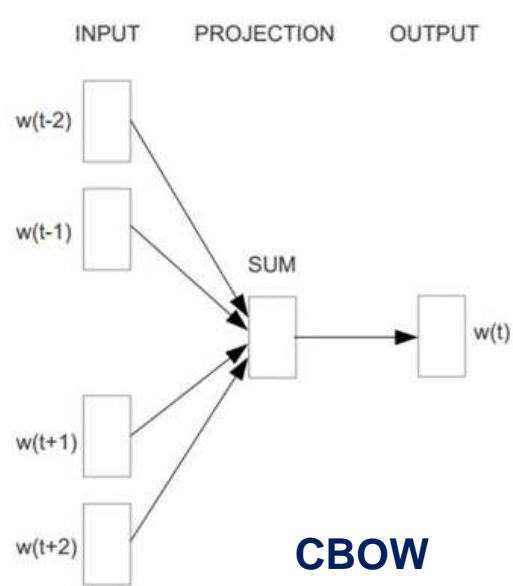
Mikolov et al, (2013a).

Semantic Similarity between Words



Kusner et al (2015)

Word2Vec



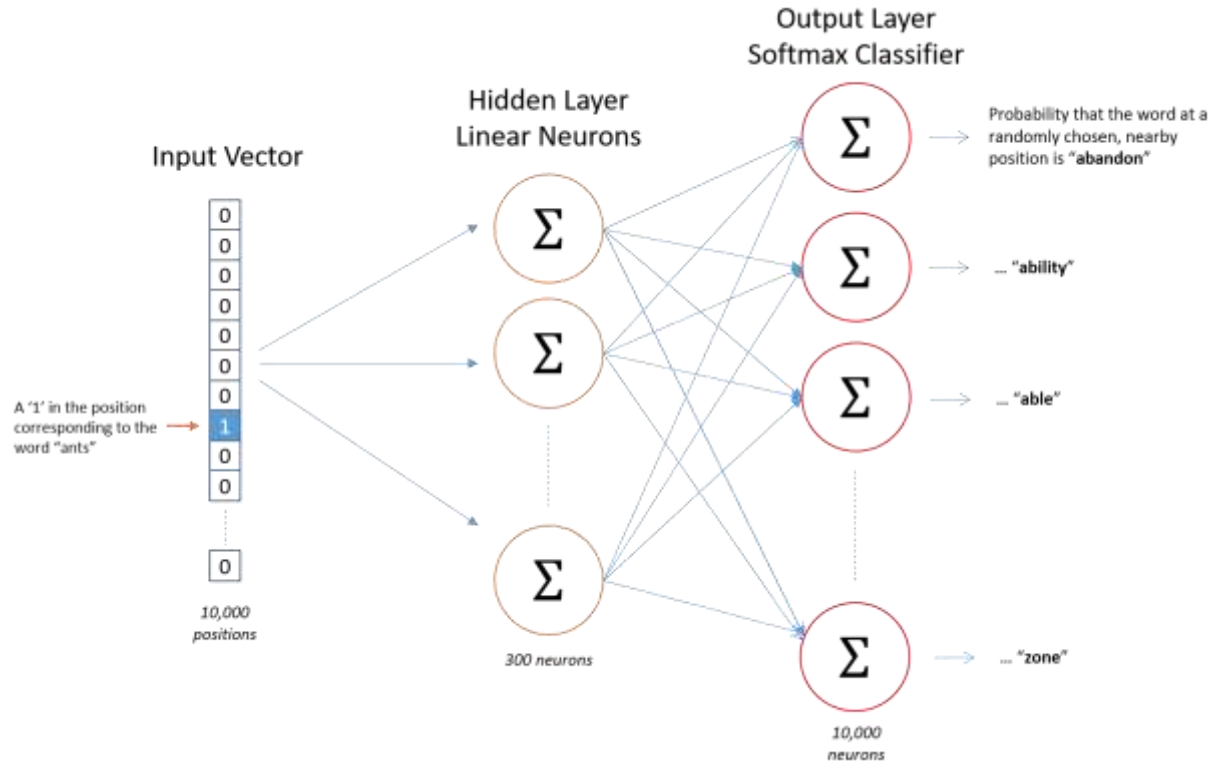
Mikolov et al, (2013a).

Word2Vec Training

Source Text	Training Samples
<div> <div>The</div> <div>quick</div> <div>brown</div> <div>fox jumps over the lazy dog.</div> </div>	<div> <div>(the, quick)</div> <div>(the, brown)</div> </div>
<div> <div>The</div> <div>quick</div> <div>brown</div> <div>fox</div> <div>jumps over the lazy dog.</div> </div>	<div> <div>(quick, the)</div> <div>(quick, brown)</div> <div>(quick, fox)</div> </div>
<div> <div>The</div> <div>quick</div> <div>brown</div> <div>fox</div> <div>jumps</div> <div>over the lazy dog.</div> </div>	<div> <div>(brown, the)</div> <div>(brown, quick)</div> <div>(brown, fox)</div> <div>(brown, jumps)</div> </div>
<div> <div>The</div> <div>quick</div> <div>brown</div> <div>fox</div> <div>jumps</div> <div>over</div> <div>the lazy dog.</div> </div>	<div> <div>(fox, quick)</div> <div>(fox, brown)</div> <div>(fox, jumps)</div> <div>(fox, over)</div> </div>

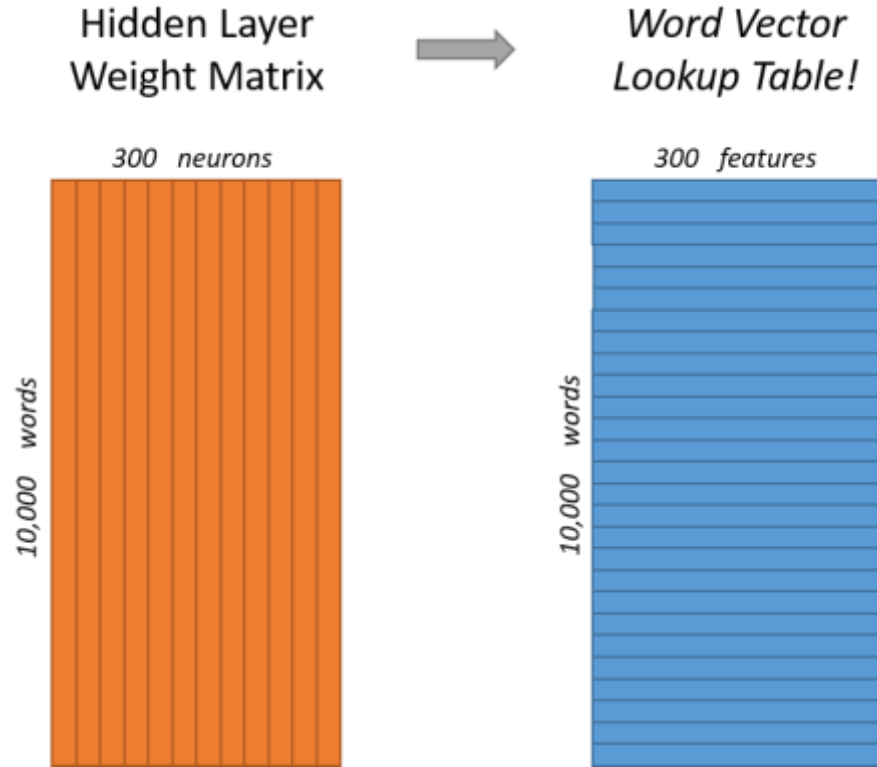
Source: <http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/>

Word2Vec Training



Source: <http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/>

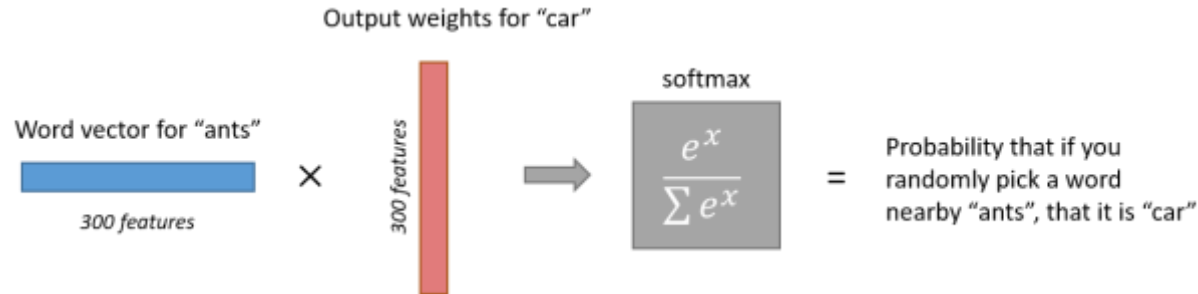
Word2Vec Training



Source: <http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/>

Word2Vec Training

$$[0 \ 0 \ 0 \ 1 \ 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \ 12 \ 19]$$



Negative Sampling

$$\log \sigma(\nu'_{w0} \cdot \nu_{wl}) + \sum_{i=1}^k \mathbb{E}_{w_i \sim P_n(w)} [\log \sigma(\nu'_{w0} \cdot \nu_{wi})]$$

$$P_n = U(s)^{3/4} / Z$$

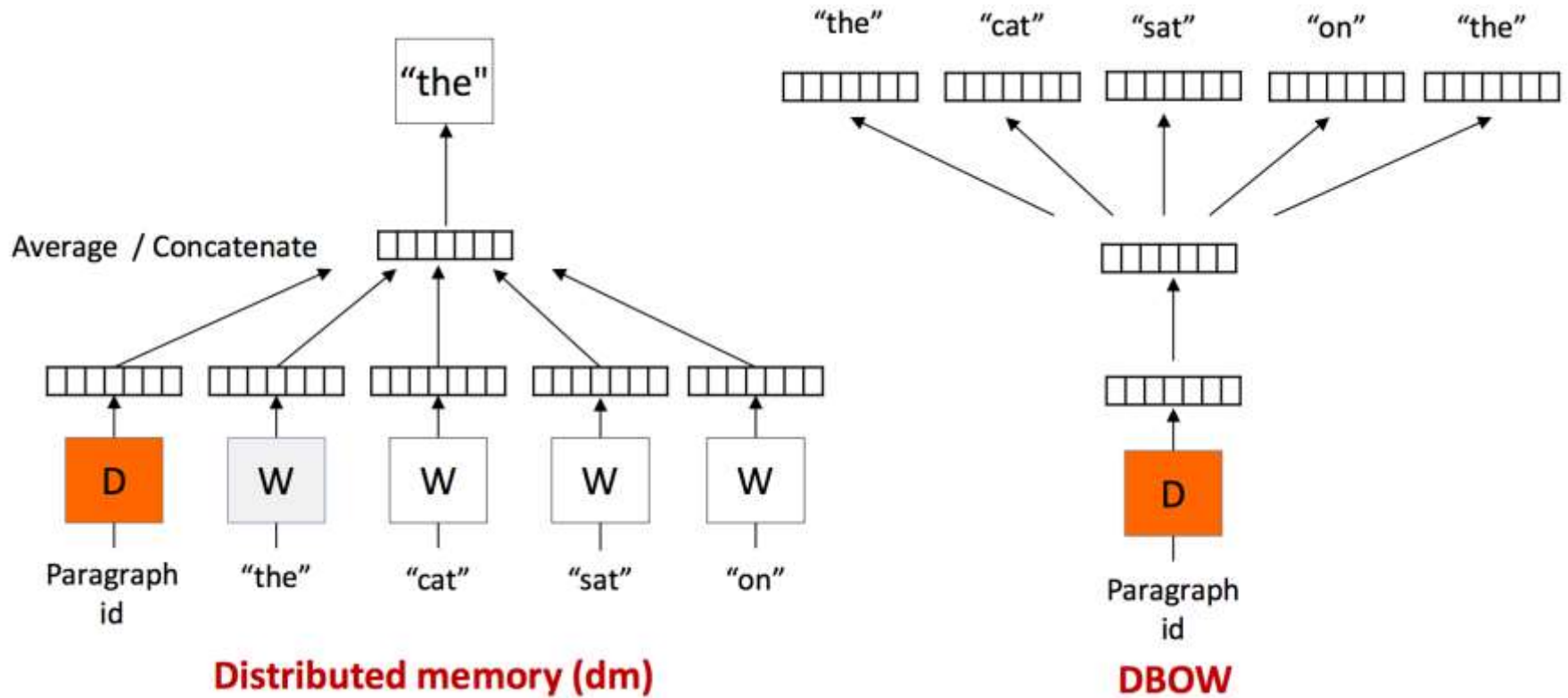
Source:

<http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/>

Fasttext

- Very Similar to Word2vec
- Primary Difference
 - Takes into account the internal structure of words while learning word representations
 - The resultant word vector is a combination of vectors for its constituent character ngrams
- Extremely fast training
- Very good for morphologically rich languages
- Takes into account both “Semantic as well as Syntactic Similarity”
- Very good performances in Syntactic Similarity tasks

Doc2Vec



Mikolov et al. (2014)

Training Parameters

Word2vec

- Skipgram
- Negative Sampling
- Dimensions = 1000
- Context Window Size = 5
- Learning Rate = 0.025
- Trained on unigrams and phrases

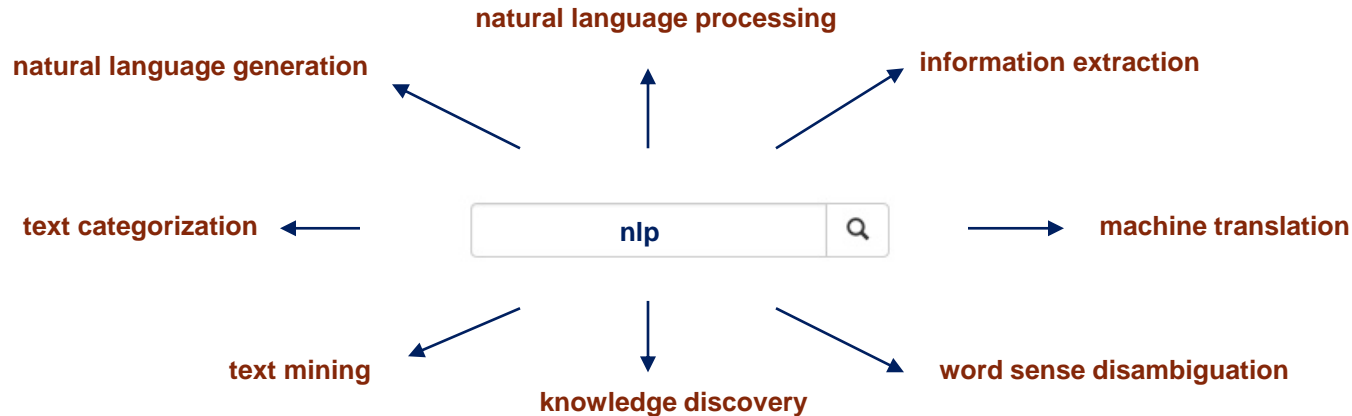
Fasttext

- Skipgram
- Negative Sampling
- Dimensions = 1000
- Max Length of Char Ngrams = 6
- Min Length of Char Ngrams = 3
- Learning Rate = 0.05
- Trained on unigrams and phrases

Doc2Vec

- Distributed Memory
- Dimensions = 1000
- Window Size = 10
- Epochs = 10
- Initial Learning Rate = 0.025
- Trained on unigrams and phrases

Query Expansion



- Getting rid of thesaurus based or dictionary based query expansion.
- How many phrases to use for expansion ?
- Which model ?
- How to integrate the trained models with a search application for query expansion ?

Recommending Similar Research Articles

From Chirps to Whistles

Discovering Event-specific Informative Content from Twitter

Debarjan Mahata
Department of Information
Science
University of Arkansas at Little
Rock
Little Rock, AR, USA
dxmahata@ualr.edu

John R. Talburt
Department of Information
Science
University of Arkansas at Little
Rock
Little Rock, AR, USA
jrtalburt@ualr.edu

Vivek Kumar Singh
Department of Computer
Science
South Asian University
New Delhi, India
vivek@cs.sau.ac.in

ABSTRACT

Twitter has brought a paradigm shift in the way we produce and consume information about real-life events. Huge volumes of user-generated tweets are produced in Twitter, related to events. Not all of these are useful and informative. A sizable amount of tweets are spurious and colloquial personal status updates, which do not provide any useful information about an event. Thus, it is necessary to identify, rank and aggregate event-specific informative content from the tweet streams. In this paper, we develop a novel generic framework based on the principle of natural metaphorism, for identifying event-specific informative content from Twitter. Mutually reinforcing relationships between tweets, hashtags, text units, URLs and users are defined and represented using *TwitterEventInfoGraph*. An algorithm - *TwitterEventInfoRank* is proposed, that simultaneously ranks tweets, hashtags, text units, URLs and users producing them, in terms of event-specific informativeness by leveraging the semantics of relationships between each of them as represented by *TwitterEventInfoGraph*. Experiments and observations are reported on four million (approx) tweets collected for five real-life events, and evaluated against popular baseline techniques showing significant improvement in performance.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]

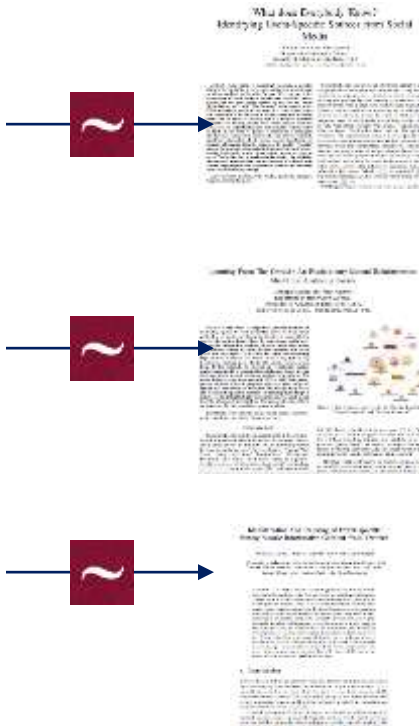
Keywords

social media mining, text mining, twitter, natural metaphorism, event, information retrieval, ranking, event-specific information

about real-life events. Twitter is one such platform that has become an indispensable source for disseminating news and real-time information about current events. It is a micro-blogging application that allows its users to post short messages of 140 characters known as tweets, from a variety of internet enabled devices. Studies have shown the importance of Twitter as a news circulation service [25], and a source for gauging public interest and opinions [18]. It's efficacy as a real-time citizen-generated source of information has been recently leveraged in detection, extraction and analysis of real-life events [26, 22, 23].

Users not only post plain textual content in their messages but also share URLs that links to other external websites, images and videos. Apart from creating new content, the users also share content produced by others. This activity is known as retweeting, and each tweet is preceded by special characters "RT". The messages are normally written by a single person and are read by many. The readers in this context are known as *followers*, and the user whose they follow is considered as their *friend*. Any user with right access either share messages that might be of interest to his followers, or for joining conversations related to the topics of his interest. The "RT" symbol followed by the microtext commonly known as user mentions, is used for mentioning other users in tweets for initiating conversations.

The explicit and inferential content of a tweet is often contextualized by the use of a cross-referenced association scheme called *hashtags*. Hashtags are a sequence of characters in any language prefixed by the symbol '#' (for e.g. #twtr4u.ac(2010)). They are widely used by the users for categorizing the content based on a topic, join conversations related to a topic, and to make the tweets easily searchable. They also act as strong identifiers of topics [13]. When tweeting about real-life events the users also tend to use *hashtags* in



Content
Similarity

Vs

Semantic
Similarity

Ranked Keyphrase Extraction

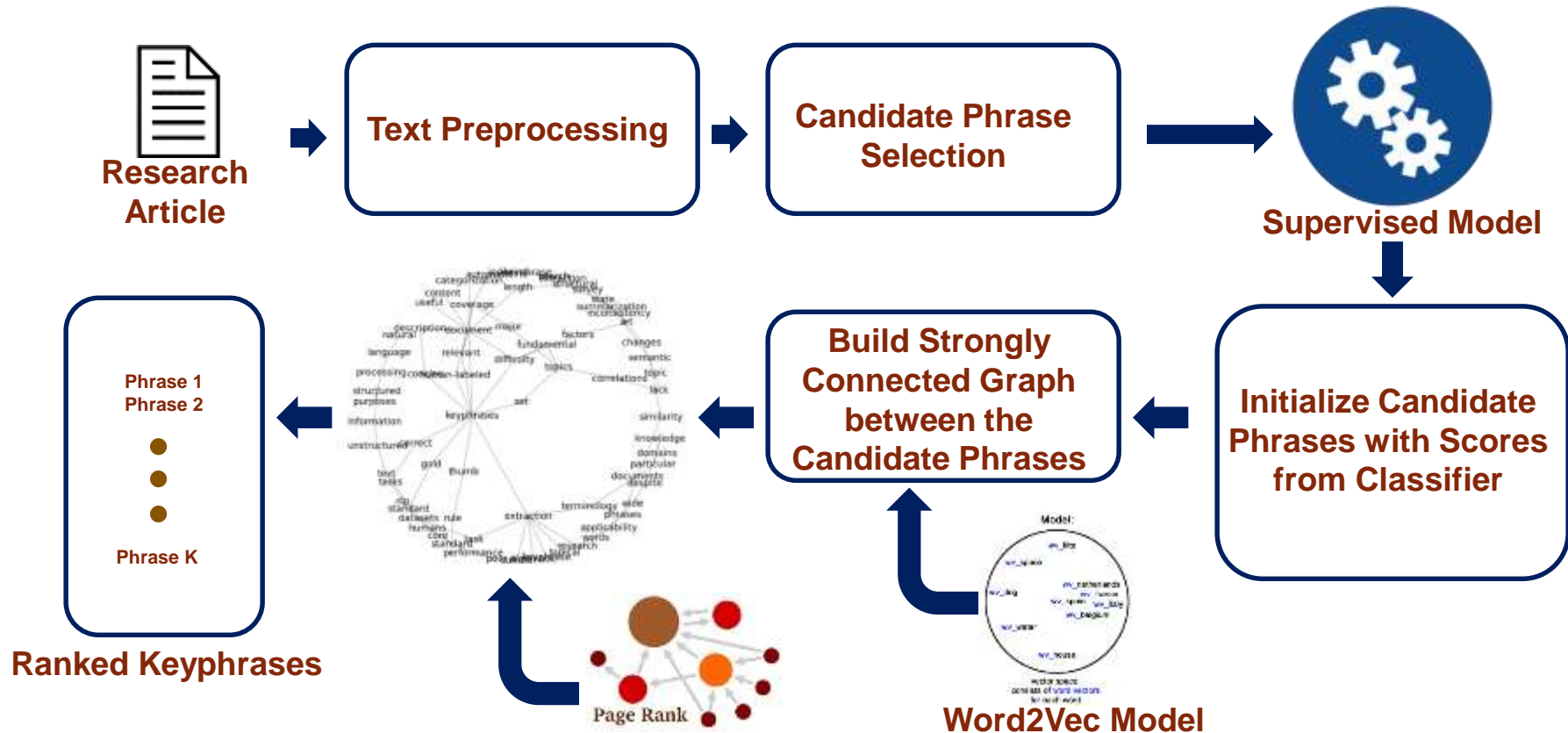
Supervised

- Known to give better results
- Drawbacks
 - Domain Specific
 - Training and Tuning of the models for generalization
 - Intelligent Feature Engineering
- Examples: KEA, MAUI

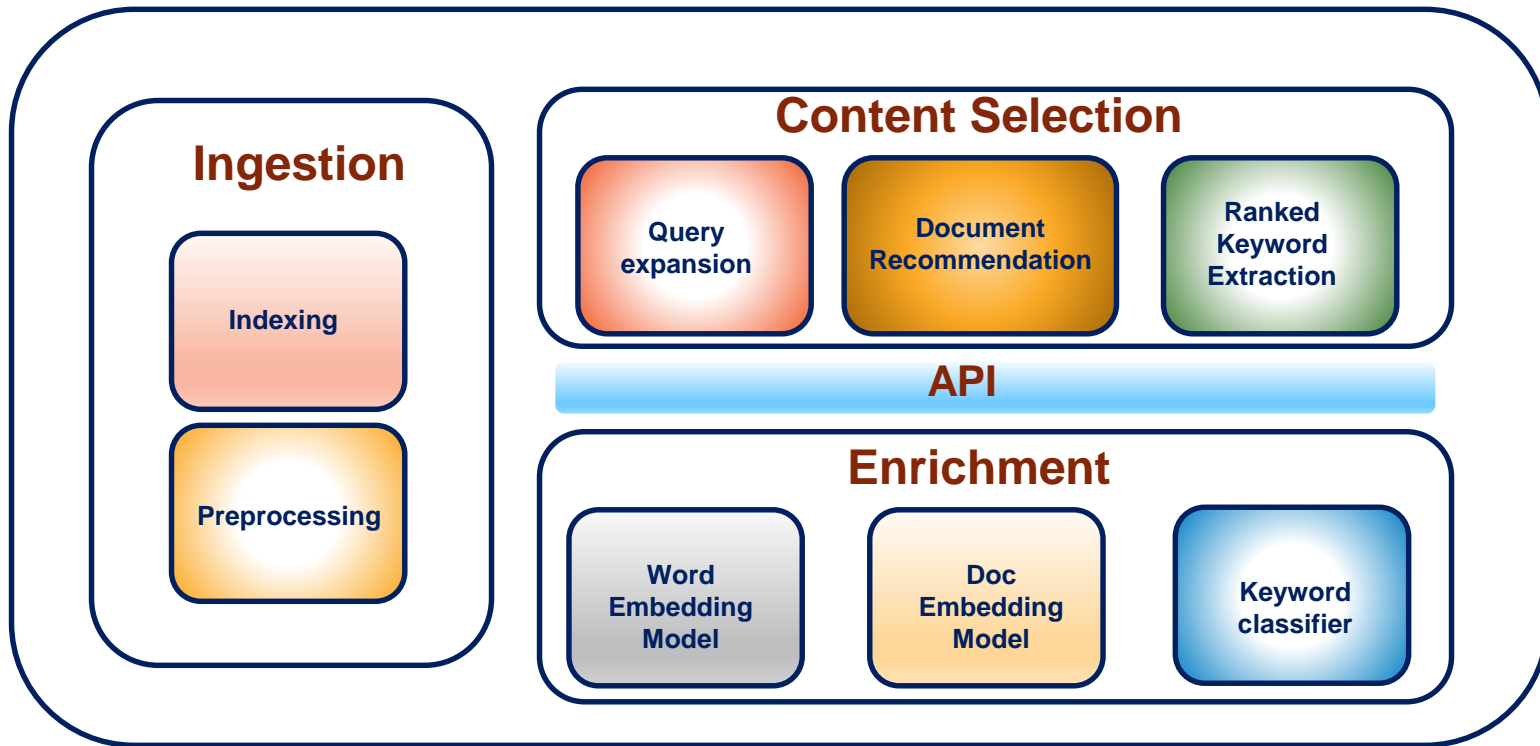
Unsupervised

- Known to give worse results than supervised in domain specific tasks
- Domain independent
- No need of feature engineering
- Algorithm determines relationships between candidates for identifying the keyphrases
- Examples: TextRank, RAKE

Ranked Keyphrase Extraction



Integration





References

1. Sebastian Rudder, “**On Word Embeddings**”; <http://sebastianruder.com/tag/word-embeddings/index.html>
2. Christopher Olah, “**Colah’s Blog**”, <http://colah.github.io/>
3. Chris McCormick, “**Word2Vec Tutorial – The Skip-Gram Model**”, <http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/>
4. Le, Quoc V., and Tomas Mikolov. “**Distributed Representations of Sentences and Documents.**” ICML. Vol. 14. 2014.
5. Mikolov, Tomas, et al. “**Distributed representations of words and phrases and their compositionality.**” Advances in neural information processing systems. 2013.
6. Bojanowski, Piotr, et al. “**Enriching word vectors with subword information.**” arXiv preprint arXiv:1607.04606 (2016).
7. Rong, Xin. “**word2vec parameter learning explained.**” arXiv preprint arXiv:1411.2738 (2014).



thank you!