
# Contextual word embeddings for recommender systems

**Manas Ranjan Kar, Practice Lead, Juxt Smart Mandate**

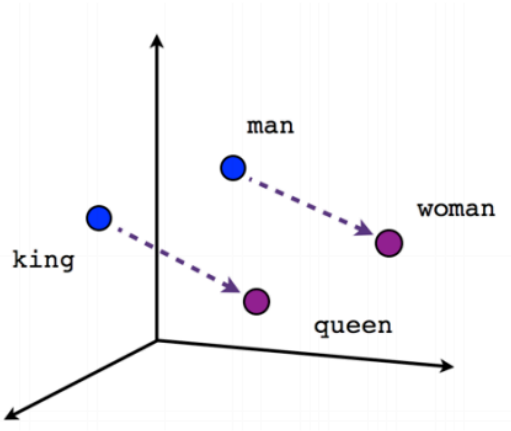
**Akhil Gupta, Intern, Amazon India**

Version 2
Date – 25.09.2016

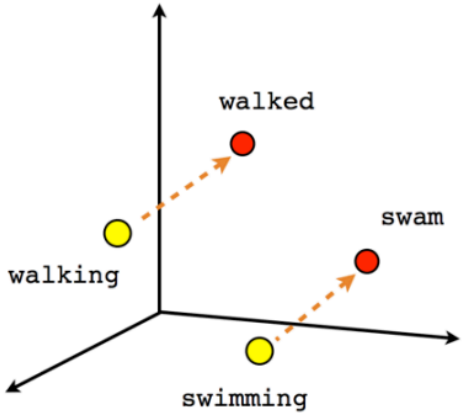
**GOAL:** To showcase a methodology to create recommendations using contextual word embeddings

***SECRET GOAL*:** To show that hypothesis/experimental tinkering can lead to real business impact + open source experiments !

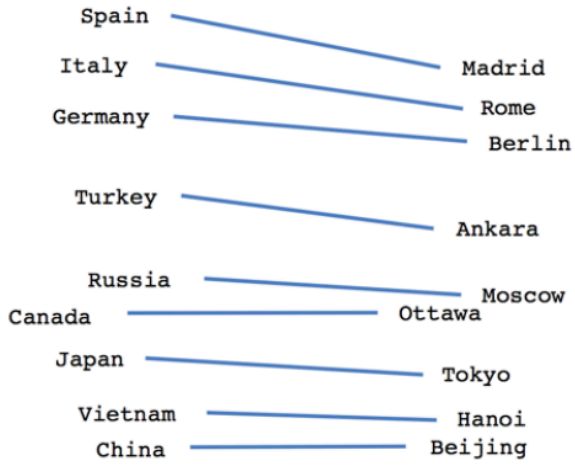
# word2vec– famous outcomes




Male-Female




Verb tense




Country-Capital

word2vec is *not* deep learning;

The whole idea behind word2vec is to demonstrate that **you can get better word representations if you trade the model's complexity for efficiency**, i.e. the ability to learn from much bigger datasets.

*“A word is known by the company it keeps”*

word2vec is a shallow word embedding model.

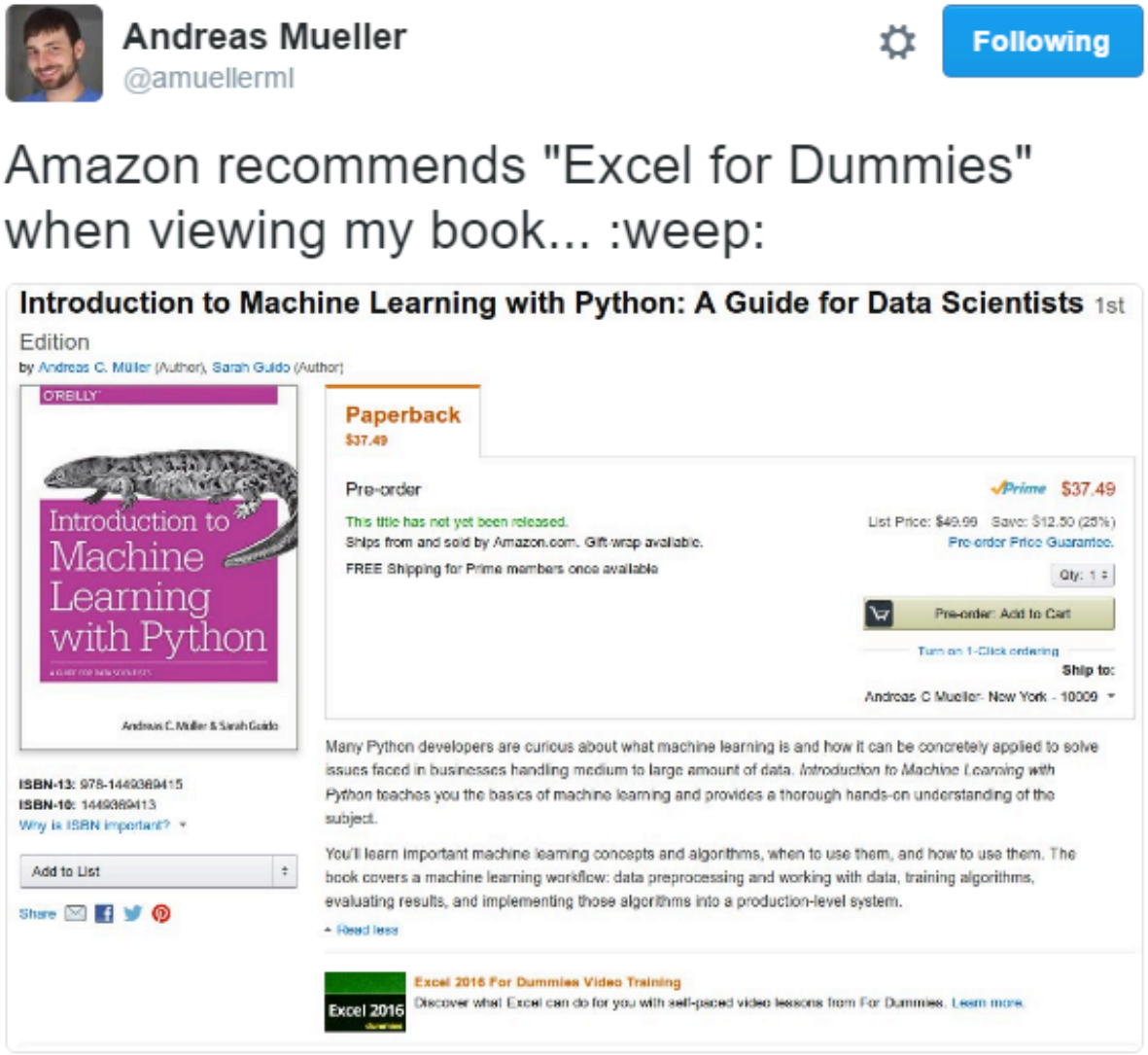
STEPS:

1. Create word IDs – 0 to length of the vocab in the corpus
2. IDs are mapped into vector space, taking their distributional properties into account
3. Training is online – one example at a time
4. Target for CBOW and SKIP respectively – predict the target word (w), given the context (c) i.e. P(w|c) OR predict the context, given the word i.e. P(c|w)

**Thumb Rule**: More data, use SKIP. Less data, use CBOW

Or not, decide depending on your experiments. Conflicting views do exist

# recommendations can be broken, well sometimes !



Word representations took the world by storm when it was introduced by Google. Numerous papers have been written after that in the domain of topic modelling, sentiment analysis and social media user profiling.

This is our attempt to propose and demonstrate a framework that’s more rounded and preserves context while generating recommendations.

Amazon Reviews and user browsing history for more than 18 years, contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.

**Digital Music  Reviews**– 800K+

**User Browsing History**: 9.4 M records

**META MODEL**:

Created with the user browsing history. This contains more than 9.4 million product histories. The attempt is to mimic and improve upon existing product-product systems.

**USER REVIEW MODEL**:

Currently many food reviews or travel sites don't allow us to search on "context". This model intakes reviews and attempts to create a framework for a "contextual search engine".

## USER REVIEWS


```
{
  "reviewerID": "A2SUAM1J3GNN3B",

  "asin": "0000013714",


  "reviewerName": "J. McDonald",
  "helpful": [2, 3],


  "reviewText": "I bought this for my
husband who plays the piano.  He is
having a wonderful time with Shankar
Mahadevan.

  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


## METADATA


```
{
  "asin": "0000031852",
  "title": "Girls Ballet Tutu Zebra
Hot Pink",
  "price": 3.17,
  "imUrl": "http://ecx.images-
amazon.com/images/I/
51fAmVkTbyL._SY300_.jpg",
  "related":
  {
    "also_bought": ["B00JHONN1S",
"B002BZX8Z6", "B00D2K1M3O", ....],
    "also_viewed": ["B002BZX8Z6",
"B00JHONN1S", "B008F0SU0Y"...],
    "bought_together": ["B002BZX8Z6"]
  },
  "salesRank": {"Toys & Games":
211836},
  "brand": "Coxlures",
  "categories": [["Sports & Outdoors",
"Other Sports", "Dance"]]
}
```




http://jmcauley.ucsd.edu/data/amazon/

**USER REVIEWS**

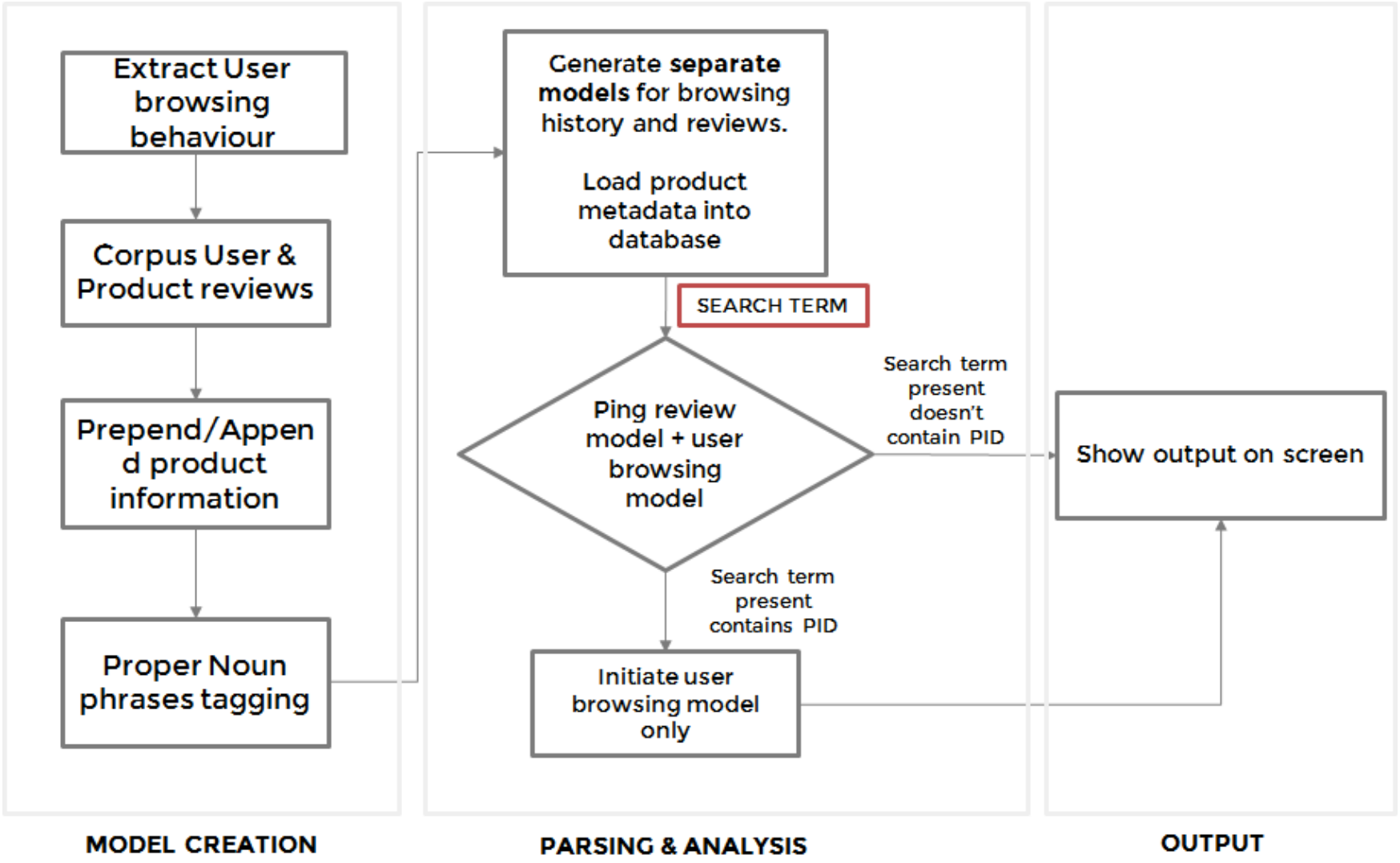
I bought this for my husband who plays the piano. He is having a wonderful time with Shankar Mahadevan.

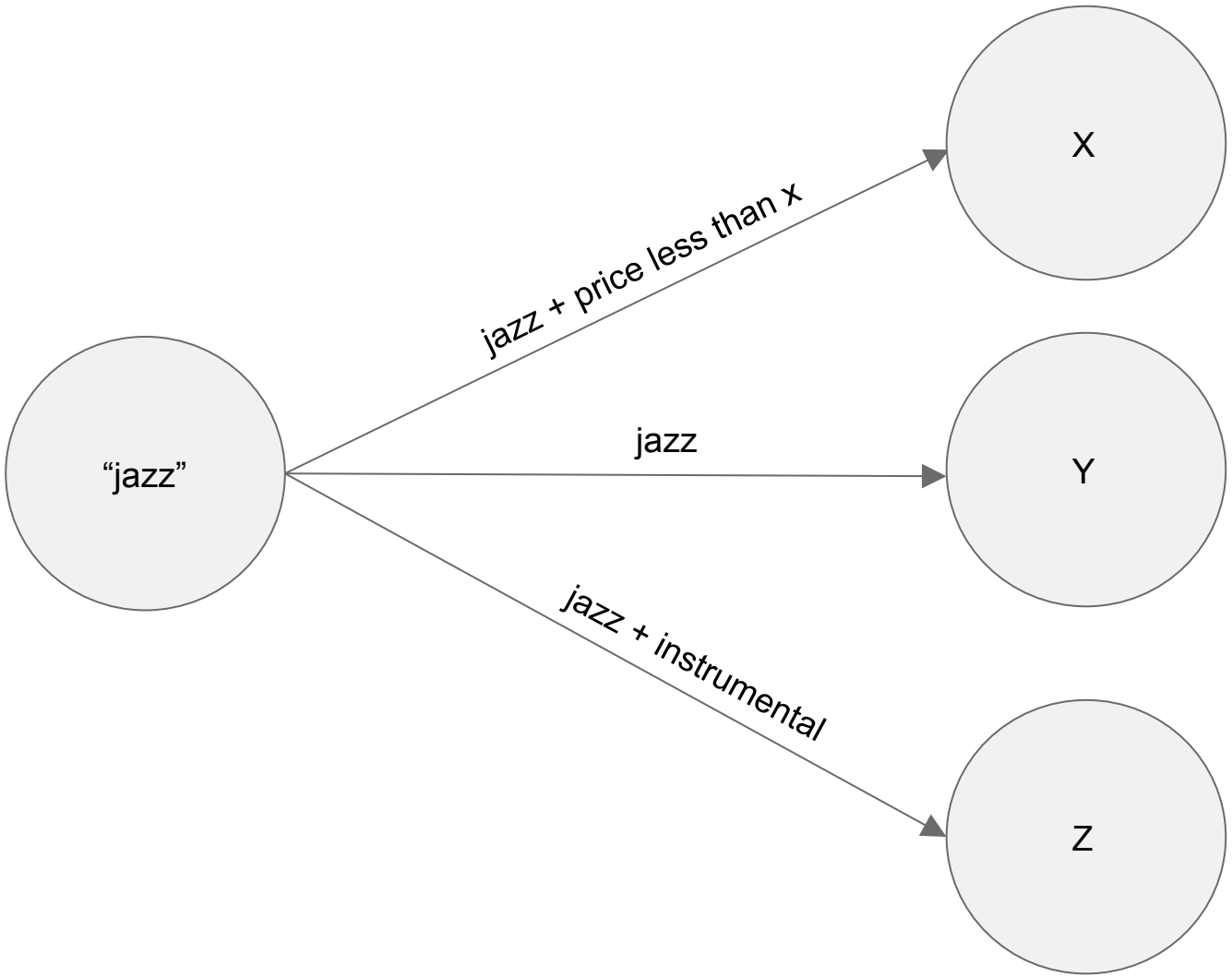
**<p_5> <0000013714>** i bought this for my husband who plays the piano. he is having a **wonderful_tim**e with **shankar_mahadevan**. **<0000013714> <p_3.17>**

**METADATA**

"0000031852", "B002BZX8Z6", “B00JHONN1S”, "B008F0SU0Y"

"0000031852", "**b_t_**B002BZX8Z6", “**a_b_**B00JHONN1S”, "**a_v_**B008F0SU0Y"


Extract User browsing behaviour
Corpus User & Product reviews
Prepend/Append product information
Proper Noun phrases tagging
Generate **separate models** for browsing history and reviews.
Load product metadata into database
SEARCH TERM
Ping review model + user browsing model
Search term present doesn't contain PID
Search term present contains PID
Initiate user browsing model only
Show output on screen
MODEL CREATION
PARSING & ANALYSIS
OUTPUT


X
jazz + price less than x
"jazz"
jazz
Y
jazz + instrumental
Z

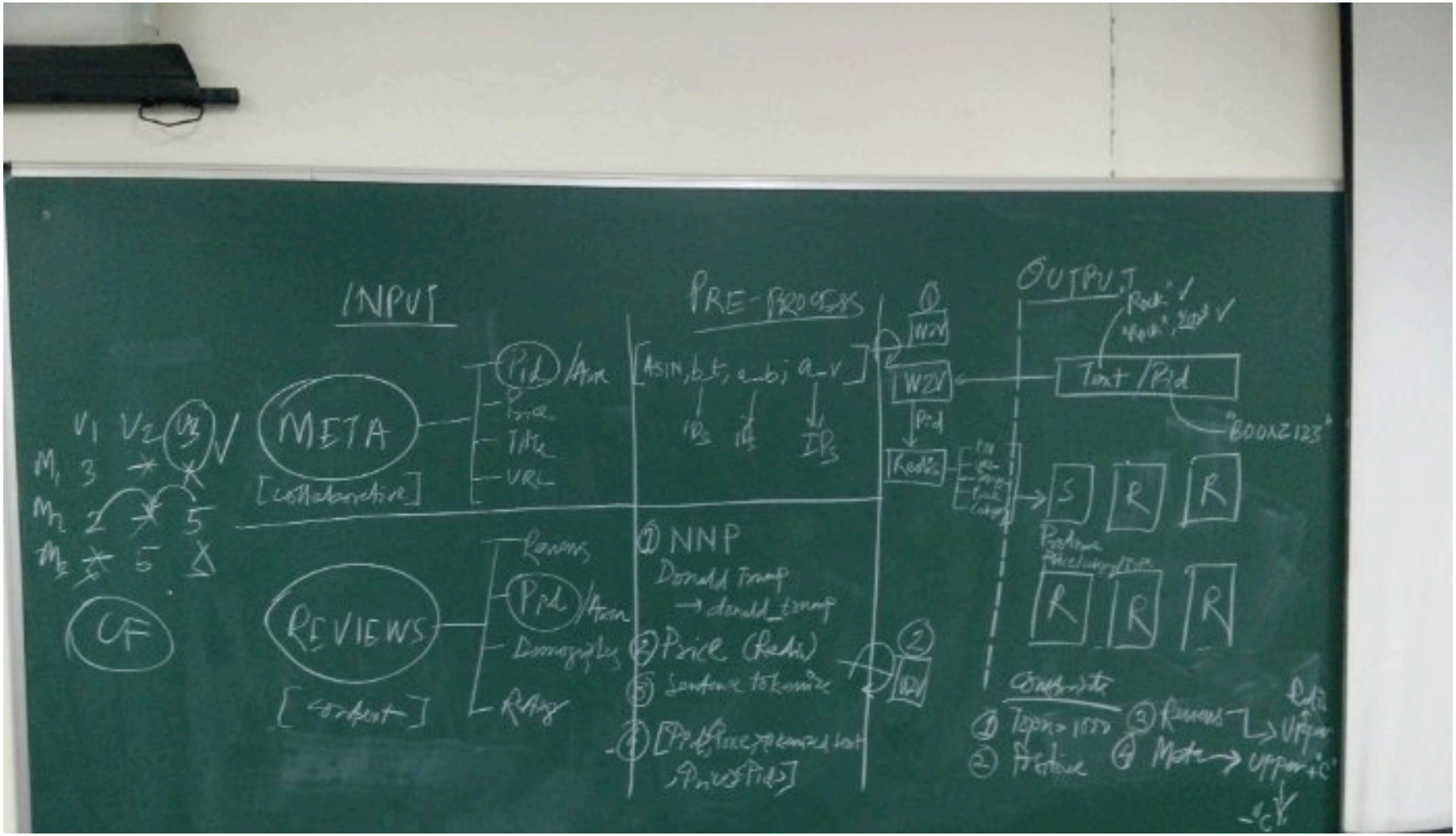
- Article/product titles re-writes for better SEO
- Compact text features during ML - **WIP**
- Validate and refine ontologies for domains
- SKU reorganization and genre validation
- Sentiment Analysis, with Sentiwordnet
- Recommendation engine for Indian restaurants

DEMO

# what could have done better

- Better pre-processing using DBpedia – currently relies on a NNP parser only
- Investigating ‘rogue’ results and experimenting with more hyper-parameters
- Mixing metadata and user reviews into one
- Developing metrics and automated techniques for measuring performance

# QUESTIONS?


INPUT
PRE-PROCESS
OUTPUT
META
[collaborative]
Pid
Asin
Price
Title
URL
REVIEWS
[content]
Reviews
Pid
Asin
Demographics
Rating
CF
W2V
Redis
Text / Pid
"B00AZ123"
S
R
R
R
R
R
NNP
Donald Trump
→ donald_trump
Price (Redis)
Sentence tokenize

https://github.com/manasRK/word2vec-recommender


| Author | Working As | contact @ |
| --- | --- | --- |
| Manas Ranjan Kar | Practice Lead @<br>Juxt Smart Mandate | manasrkar91@gmail.com |
| Akhil Gupta | Intern, Amazon India | akhilgupta.official@gmail.com |