Understanding **Product Attributes** from Reviews.

Sanghamitra Deb
Accenture Technology Laboratory

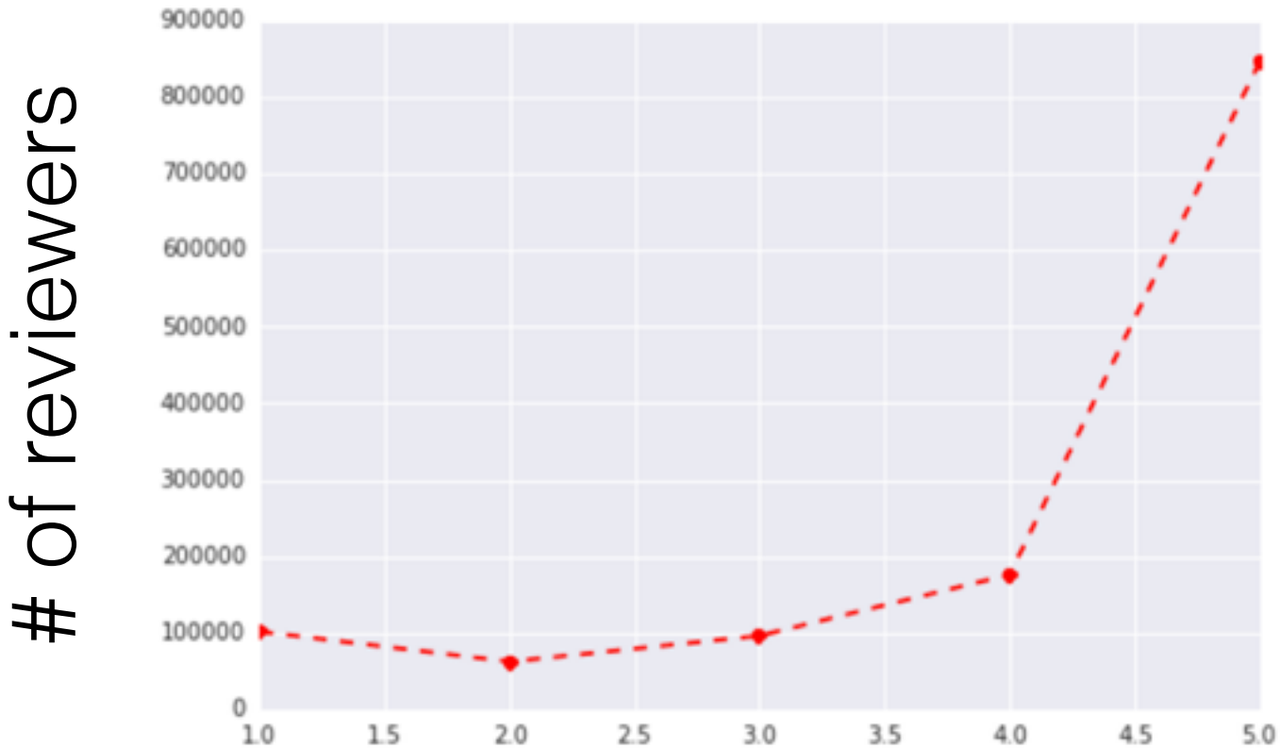
Why Product Attributes?

- Find **darjeeling tea** that **ships within two days**?
- Which **power bars** are **not too sweet**?
- Which **wine** and **cheese** should be marketed together?
- Does the **icing** have added **color**?
- Is the **soy milk GMO free**?

Products: Amazon Gourmet-Food

Public Data: Amazon

{u'asin': u'B00LQWKDBM',
 u'helpful': [0, 0],
 u'**overall**': 5.0,
 u'**reviewText**': u"Makes AMAZING meatloaf. It's our "old family recipe"!",
 u'reviewTime': u'07 12, 2014',
 u'reviewerID': u'A2NSZZ7Y0RAE45',
 u'reviewerName': u'Michael Freed "freedml"',
 u'summary': u'AMAZING meatloaf',
 u'unixReviewTime': 1405123200}



Reviews

reviews:
"No sugar, no GMO garbage, no fillers that come with store bought extracts. This stuff is just amazing. I use it in everything from baking to cooking and even as suggested in my coffee which is saying a lot because I normally do not care for flavored coffee! You cannot go wrong with this. I've ordered from this merchant before, customer satisfaction is their priority and service was quick, shipped right out with tracking even! I'll be buying from GLS Goods again! I won't use any other vanilla!" .

“Wow. This stuff is hot. As expected. I truly love the smoky flavor. This jar will last a while, thus the slightly higher price is OK.”

"If you've got a craving for something different this is for you. Same old Kit Kat wafer but the coating is white chocolate flavored with green tea. Really unique. But if you have a raging sweet tooth you probably won't like it, it's not too sweet.”

“I ordered 3 types of Power Bar. None of them tastes fresh. The expiration date isn't until September (it's June now) but I think these bars are stale. I won't be reordering.”

“This brand is disgusting. It is watery and too sweet. If you like the Tazo brand of chai latte you will HATE this brand. I returned it.”

Reviews

reviews:
"**No sugar**, **no GMO** garbage, no fillers that come with store bought extracts. This stuff is just amazing. I use it in everything from **baking** to **cooking** and even as suggested in my coffee which is saying a lot because I normally do not care for flavored coffee! You cannot go wrong with this. I've ordered from this merchant before, **customer satisfaction** is their priority and service was quick, shipped right out with tracking even! I'll be buying from GLS Goods again! I won't use any other vanilla!" .

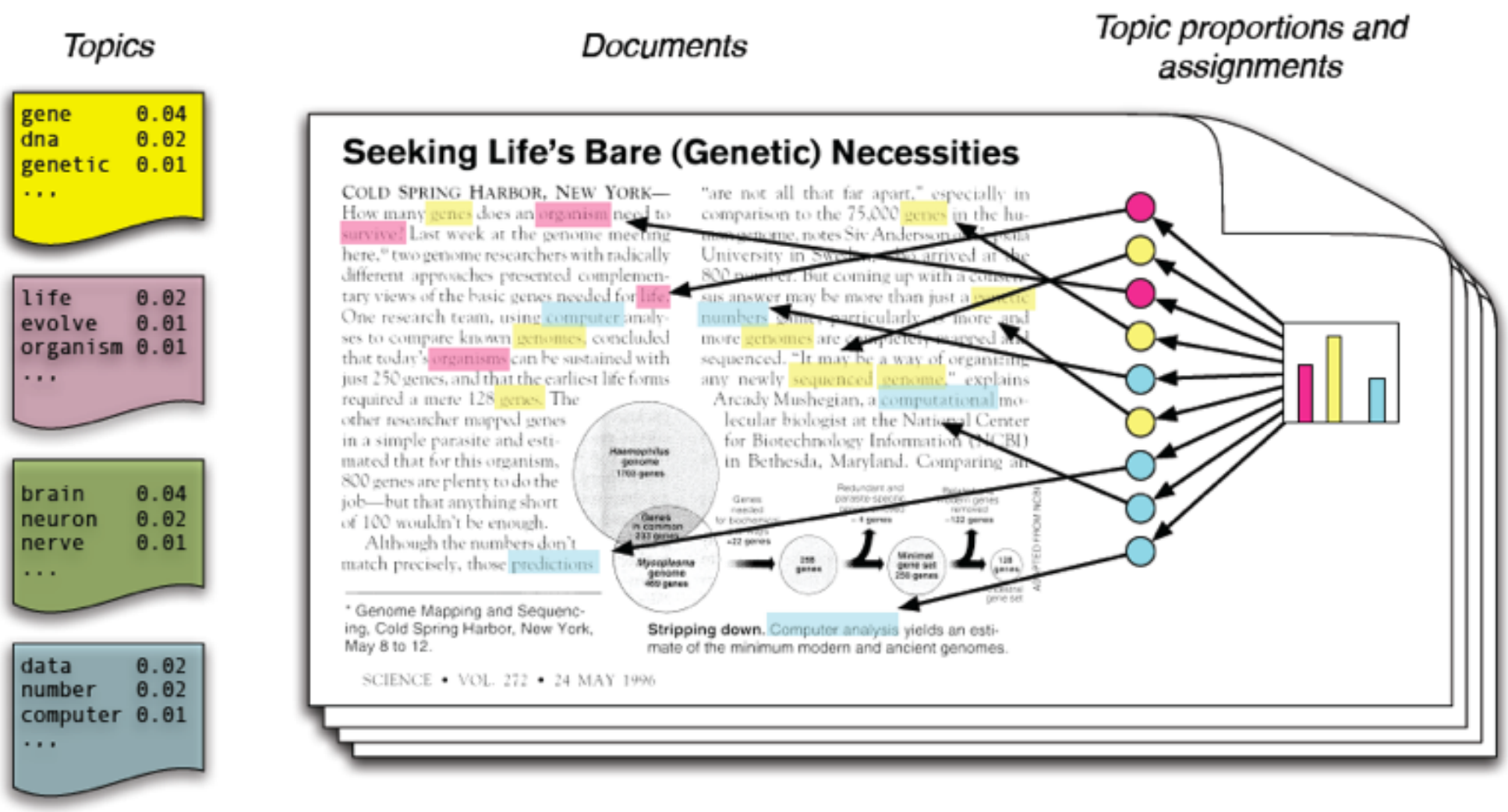
“Wow. This stuff is **hot**. As expected. I truly love the **smoky** flavor. This jar will last a while, thus the slightly higher price is OK.”

"If you've got a craving for something different this is for you. Same old Kit Kat wafer but the coating is **white chocolate** flavored with **green tea**. Really unique. But if you have a raging sweet tooth you probably won't like it, it's **not too sweet**.”

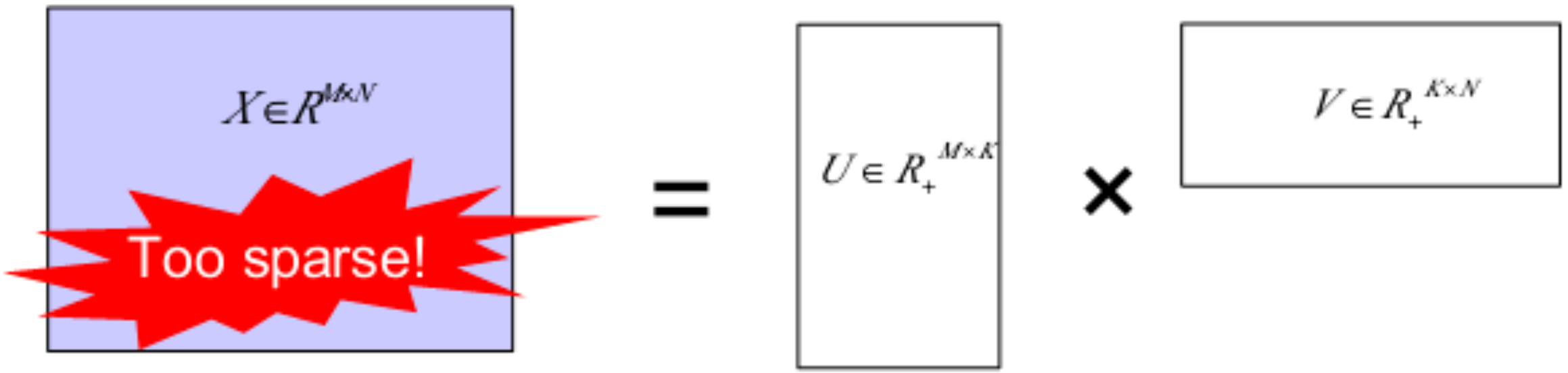
“I ordered 3 types of Power Bar. **None** of them **tastes fresh**. The expiration date isn't until September (it's June now) but I think these bars are **stale**. I won't be reordering.”

“This brand is **disgusting**. It is watery and **too sweet**. If you like the Tazo brand of chai latte you will **HATE** this brand. I returned it.”

Product Attributes: Topic Modeling



Topic Modeling using Non-negative Matrix Factorization



term-doc matrix

term-topic matrix

topic-doc matrix

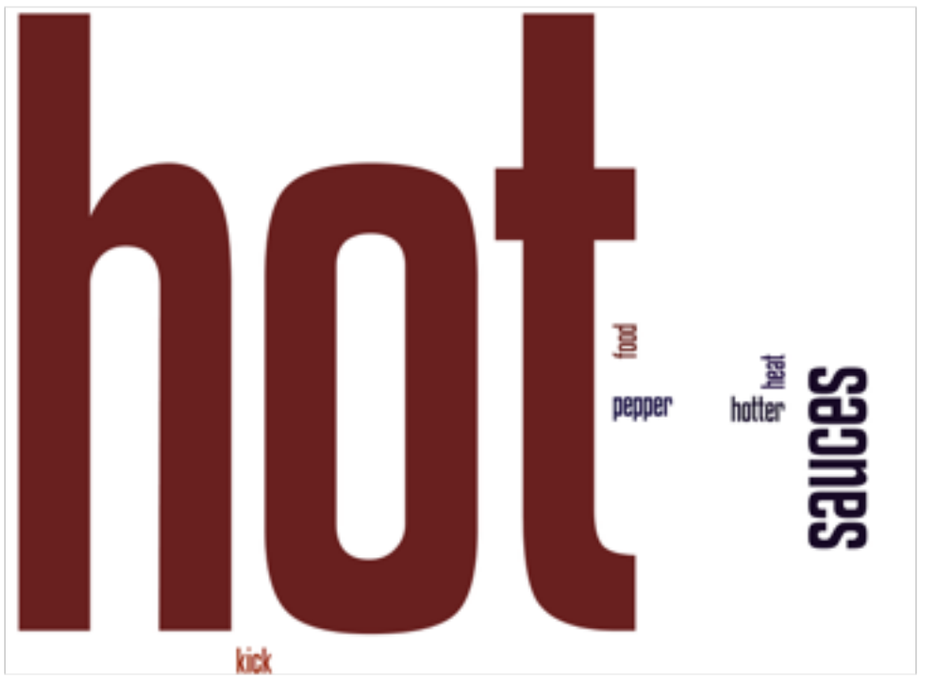

salad
salads
oils
olives
extra
cooking
bread
add
butter
stir
wonderful
peanut
vinegar
delicious
fry
italy
italian
garlic
olive
oil


shipping
product
received
food


cheese
assortment
crackers
cheddar
creamy
wonderful
wine
aged
goat
blue
cheeses


cup
orange
tazo
loose
harney
enjoy
earl
black
bag
drink
leaves
green
iced
drinking
grey
blend
tea
bags
strong


tea
bags
loose
earl
grey
teas
tazo
black
drink


hot
food
pepper
heat
hotter
sauces
kick


fresh
arrived
packaged
fast
fruit
quickly
jelly


tree
came
soil
water
leaves
plant
bonsai
care
condition
looks
trees
packed
arrived
healthy


coffee
chicory
filter
cup
water
drip
maker
pot

Topic Modeling : Finding **coherent** set of words that occur together.

(1) Identifying topics: Maximum weight
(2) Using topic words such as "**shipping**", "**tea**", "**Darjeeling**" to identify attributes

(1) Find the product reviews with topic words such as "**shipping**", "**tea**", "**Darjeeling**" to identify attributes
(2) Develop a **sentiment analyzer** from **star ratings** on the entire review data set.
(3) Use the **sentiment analyzer** to predict sentiment of sentences with attributes.

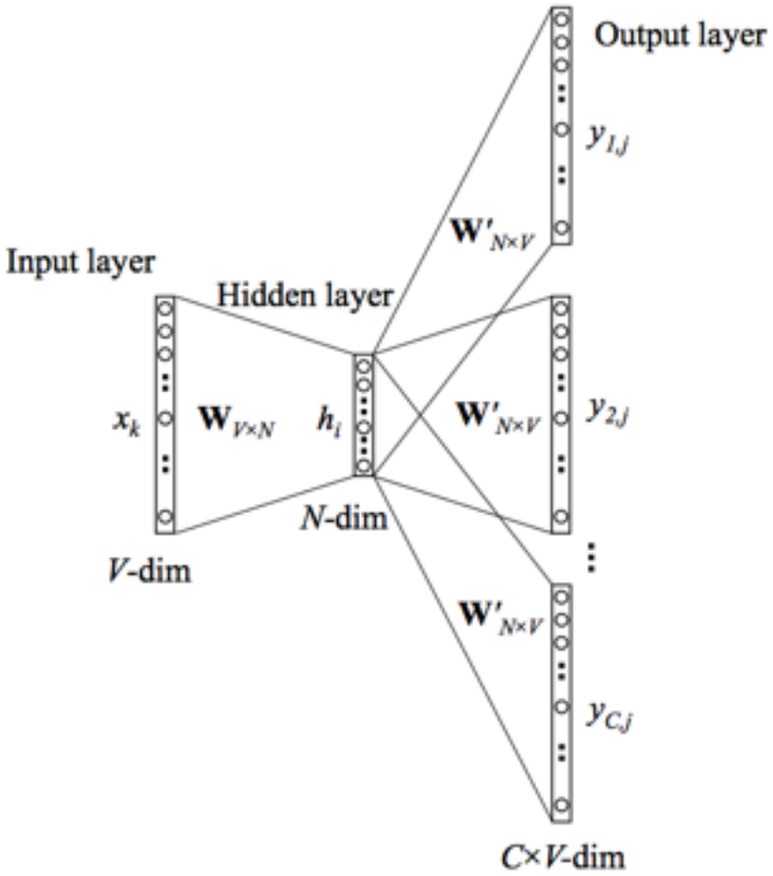
Attribute Sentiments

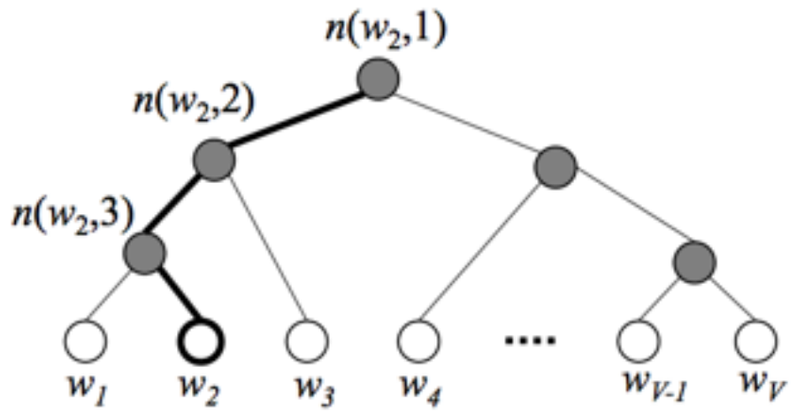
'This **chili** powder adds a **wonderful** spiciness and smokiness to dishes. Warning, however. **Ghost** peppers have a spiciness that increases after the first taste, so adjust accordingly.'

"No sugar, no GMO garbage, no fillers that come with store **bought** extracts. This stuff is just amazing. I use it in everything from **baking** to **cooking** and even as suggested in my coffee which is saying a lot because I normally do not care for flavored coffee! You cannot go wrong with this. I've ordered from this merchant before, customer satisfaction is their priority and service was quick, shipped right out with tracking even! I'll be buying from GLS Goods again! I won't use any other **vanilla**!"

Demo

Undersatnding Products:Word2Vec






```
import logging
logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s',\
    level=logging.INFO)

num_features = 300    # Word vector dimensionality
min_word_count = 40   # Minimum word count
num_workers = 4       # Number of threads to run in parallel
context = 10          # Context window size
downsampling = 1e-3   # Downsample setting for frequent words

# Initialize and train the model (this will take some time)
from gensim.models import word2vec
print "Training model..."
model = word2vec.Word2Vec(all_text,
                          workers=num_workers,size=num_features,
                          min_count = min_word_count, window = context, sample = downsampling)

# If you don't plan to train the model any further, calling
# init_sims will make the model much more memory-efficient.
model.init_sims(replace=True)

# It can be helpful to create a meaningful model name and
# save the model for later use. You can load it later using Word2Vec.load()
model_name = "300features_40minwords_10context"
model.save(model_name)
```

```
Training model...
```

Tea

blossoms
teas
moroccan
scented
lemongrass
hibiscus
hips
petals
chamomile
ceylon
india
bud
herbal
harney
spearmint
mist
zen
stash
jasmine
lotus
oolong
blossom
blackberry
bush
ginseng
flowery
balm
bergamot
bigelow

Coffee

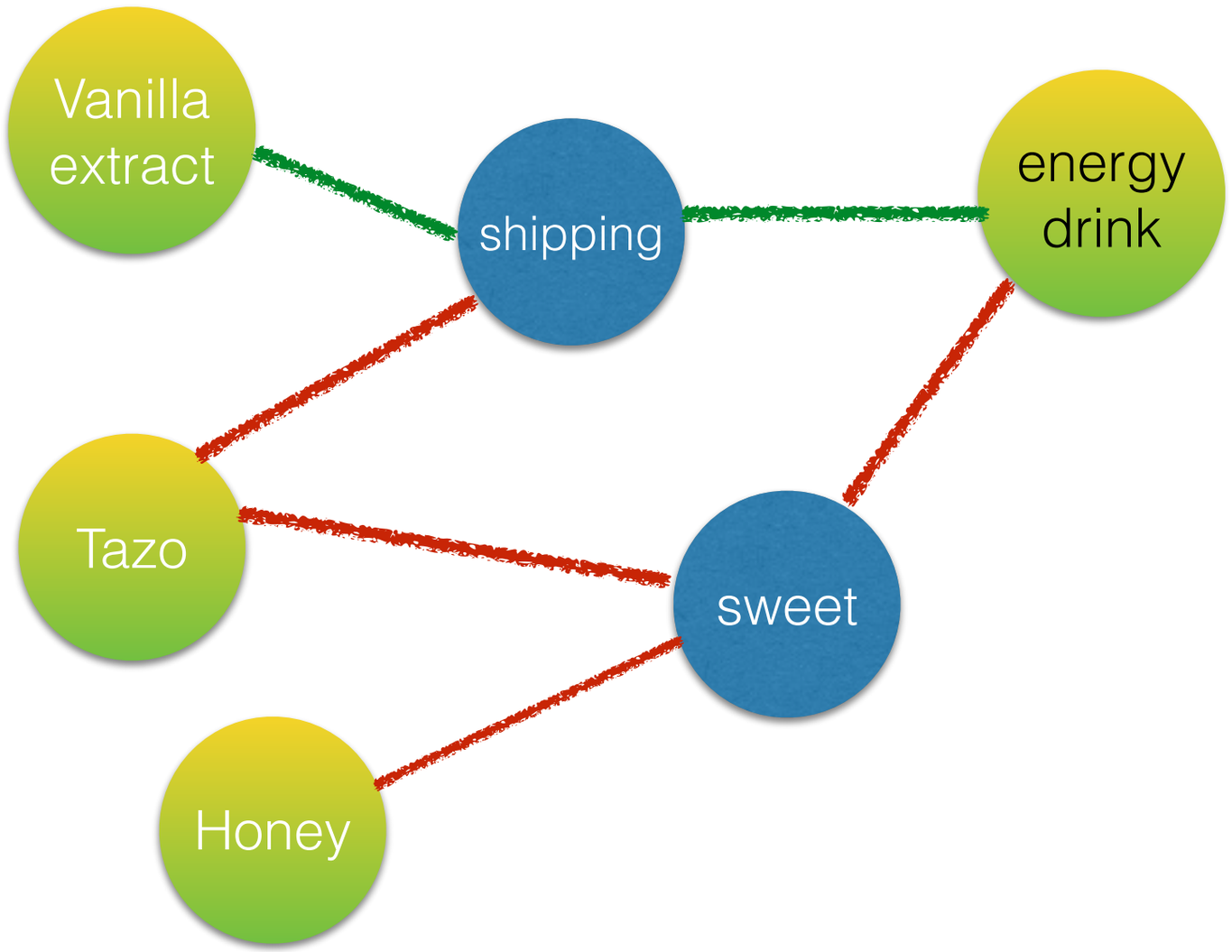
espresso
machine
cappucino
turkish
maker
machines
grind
cappuccino
coffees
expresso
pod
grinding
lavazza
illy
press
melitta
french
java
cafe
pods
keurig
senseo
roast
crema
hamilton
bunn
makers
holder
drip
frothy
kona
brewer
folgers
ese
model
britt
cleanup
gran
button
aloha
beach
roasts
grounds
cones

wow this arrived with two green rings the same size and no yellow ring where is the quality control didn t the packager at amazon notice this glaring mistake what a disappointment feels cheap and smells bad pass on this one **B00000IRTW** didn t the packager at amazon notice this glaring mistake what a disappointment feels cheap and smells bad pass on this one

Product id

Demo

Knowledge graph



Extracting relationships

DeepDive:
Quantitative information on creating structure from unstructured text
You have to know the relationships you are trying to extract and the data model for your problem.

define relationship: **is_tea/coffee_maker**
find a training set corresponding “tea_maker”, i.e if the sentences contain these keywords “**melitta**”,”**bodum**”, …
Identify features of the data: **Organization names**

http://deepdive.stanford.edu/

Parsed Data input to deepdive


```
0      1      No sugar, no GMO garbage, no fillers that come with store bought extracts.
"{""No"",""sugar"","","",""no"",""GMO"",""garbage"","","",""no"",""fillers"",""that"",""come"",""with"","
"store"",""bought"",""extracts"","".""}"

"{""no"",""sugar"","","",""no"",""GMO"",""garbage"","","",""no"",""filler"",""that"",""come"",""with"",""
store"",""buy"",""extract"","".""}"
"{""DT"",""NN"","","",""DT"",""NNP"",""NN"","","",""DT"",""NNS"",""WDT"",""VBP"",""IN"",""NN"",""VBD"",""
NNS"","".""}"
"{""O"",""O"",""O"",""O"",""ORGANIZATION"",""O"",""O"",""O"",""O"",""O"",""O"",""O"",""O"",""O"",""O"",""
O""}"  {0,3,8,10,13,17,24,26,29,37,42,47,52,58,65,73}
"{""neg"","""","""",""neg"",""nn"",""appos"","""","",""neg"",""appos"",""nsubj"",""rcmod"",""mark"",""nsubj"
","advcl"",""dobj"","""}"    {2,0,0,6,6,2,0,9,2,11,9,14,14,11,14,0}
```


Training set:

Positive

Mellitta	“tea/coffee maker”
bodum	“tea/coffee maker”

Training set:

Negative

Cuisinart	“not_tea/coffee maker”
Vitamix	“tea/coffee maker”

Output: SQL tables with exception (0,1) of whether a review is a tea/coffee maker

public	is_tea	table	sanghamitra_deb	4784 kB
public	is_tea_features	table	sanghamitra_deb	52 MB
public	is_tea_is_true_calibration	view	sanghamitra_deb	0 bytes
public	is_tea_is_true_inference	view	sanghamitra_deb	0 bytes
public	is_tea_is_true_inference_bucketed	view	sanghamitra_deb	0 bytes
public	sentences_reviews	table	sanghamitra_deb	69 MB

Future Work: Text + Images

No of. products with 1 review:
66743
and 1500 reviews<20 words:

example: "I just purchased this item, and I'm happy with my purchase. I think the flavor is great."

"too small to work in an air popper. Who uses the stove anymore!!!"

Thank You.
@sangha_deb
deblivingdata.net,
sangha123.github.io
sangha123/PyData2015NYC

Data Discovery