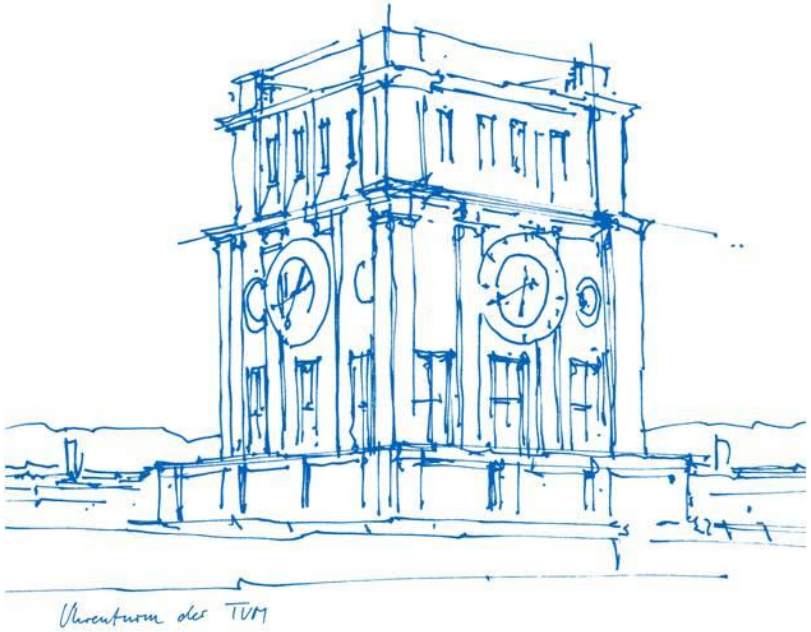
Multi-Lingual Theme Prediction of Customer Reviews Using Deep Pre-Trained Embeddings

Team 06

Michael Sorg

08.05.2019



Amazon Multilingual Dataset

(678993, 15)

	marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating	helpful_votes	total_votes	vine	verified_purchase	review_headline	review_body	review_date
0	DE	10133	RVOG49N0H1FB6	B004TACMZ8	569741360	Bosch GMS120 Ortungsgerät digital multi-Scanner	Home Improvement	5	0	0	N	Y	Super	Delivery took a little bit more then i expecte...	2014-08-01
1	DE	19612	RNCMD6OLTP4HM	1846071224	785505948	The Wheels On The Bus: Favourite Nursery Rhyme...	Books	5	1	1	N	Y	Great compilation	We enjoy listening to the song as preparation ...	2014-12-04
2	DE	19612	R4AUOBI8YC0R8	0375851569	516548029	Dr. Seuss's Beginner Book Collection	Books	5	0	0	N	Y	Great Collection	Very great compilation. Interesting story and ...	2014-12-04
3	DE	19677	R1VSHIJ1RHIBTE	B0060SVG54	302116447	Zwei an einem Tag	Video DVD	5	0	0	N	Y	Guter Verfilmung	Den Film habe ich bereits vor lesen des Buches...	2015-07-16
4	DE	19999	R3JBLVALWSLCZD	B00EYQ6CVC	368843515	Dr. House - Die komplette Serie, Season 1-8 (L...	Video DVD	5	9	14	N	Y	Kauft diese Box!	Die Box ist super verarbeitet, sieht gut aus b...	2014-02-08

```
ar: 1
fa: 1
zh-cn: 1
bg: 2
ko: 2
lv: 2
tr: 2
cs: 4
lt: 5
ro: 8
vi: 8
tl: 15
fi: 17
ru: 17
sl: 18
hu: 21
pl: 21
sk: 23
hr: 24
id: 24
et: 32
sq: 33
pt: 54
no: 81
sv: 81
so: 82
da: 105
cy: 106
af: 108
it: 150
ca: 158
nl: 179
es: 180
fr: 307
en: 48660
de: 628461
```

Review

- For each dataset create tensorflow dataset loader (using dataset api)
- Amazon Reviews dataset does not fit on disk and into ram → change runtime in colab

```
[75]  l = ["cache/reviews_Arts_Crafts_and_Sewing_5.json.gz",
           "cache/reviews_Baby_Products_5.json.gz"]
```

```
[86]  d = tf.data.TextLineDataset(l, compression_type="GZIP")
      d = d.shuffle(1024).batch(16).prefetch(16)
```

```
iterator = d.make_one_shot_iterator()
next_element = iterator.get_next()

with tf.Session() as sess:
    for i in range(6):
      value = sess.run(next_element)
      print(value)
```

```
[b'{"reviewerID": "A3TYDOH5JLRD37", "asin": "B000BNLLHW",
 b'{"reviewerID": "A1LV9A437V9X6K", "asin": "B000OMZXGU",
 b'{"reviewerID": "A2QGLMPBVZ3OYY", "asin": "B000AM7YJI",
 b'{"reviewerID": "A35G9GHVA9WHD4", "asin": "B0007XMDIM",
```

Review

☑ Full training pipeline with dummy architecture

Roadmap

1. network architecture search (train/test on amazon multilingual en/de)
   - w/o training xling embeddings

2. Additionally use full amazon reviews dataset

3. Test and fine-tune on organic dataset

Questions

- Docker instances → Google account required

(568454, 10)

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

De: 600k, UK: 1,7m, US: 7M

(678993, 15)

	marketplace	customer_id	review_id	product_id	product_parent	product_title	product_category	star_rating	helpful_votes	total_votes	vine	verified_purchase	review_headline	review_body	review_date
0	DE	10133	RVOG49N0H1FB6	B004TACMZ8	569741360	Bosch GMS120 Ortungsgerät digital multi-Scanner	Home Improvement	5	0	0	N	Y	Super	Delivery took a little bit more then i expecte...	2014-08-01
1	DE	19612	RNCMD6OLTP4HM	1846071224	785505948	The Wheels On The Bus: Favourite Nursery Rhyme...	Books	5	1	1	N	Y	Great compilation	We enjoy listening to the song as preparation ...	2014-12-04

(8823, 12)

	Author_ID	Author_name	Comment_number	Sentence_number	Domain_Relevance	Sentiment	Entity	Attribute	Sentence	Source_file	Annotator	Aspect
0	Justin-Ma	Justin Ma	521	1	0	NaN	NaN	NaN	Thanks for the thoughtful response.	quora.json	sumit	nan-nan
1	Justin-Ma	Justin Ma	521	2	0	NaN	NaN	NaN	I think we actually have a lot of common groun...	quora.json	sumit	nan-nan
2	Justin-Ma	Justin Ma	521	3	0	NaN	NaN	NaN	All I want to emphasize are my main points: Pr...	quora.json	sumit	nan-nan
3	Justin-Ma	Justin Ma	521	4	9	p	cg	pp	Industrialization is everything about producti...	quora.json	sumit	cg-pp