بسم الله الرحمن الرحیم

پروژه غنی سازی متن

ارایه‌دهندگان:
مهدی سعیدی-محمدرضا کمالی- فرانک کریمی - آرمان مظلوم زاده

استاد درس:
دکتر احسان الدین عسگری

جمع آوری دیتا

- حکمت های نهج البلاغه
- دیوان شمس
- مجموعه غزل های حافظ
- مجموعه تک بیتی از دیگر شاعران زبان فارسی
- مجموعه از نقل و قول های زبان انگلیسی

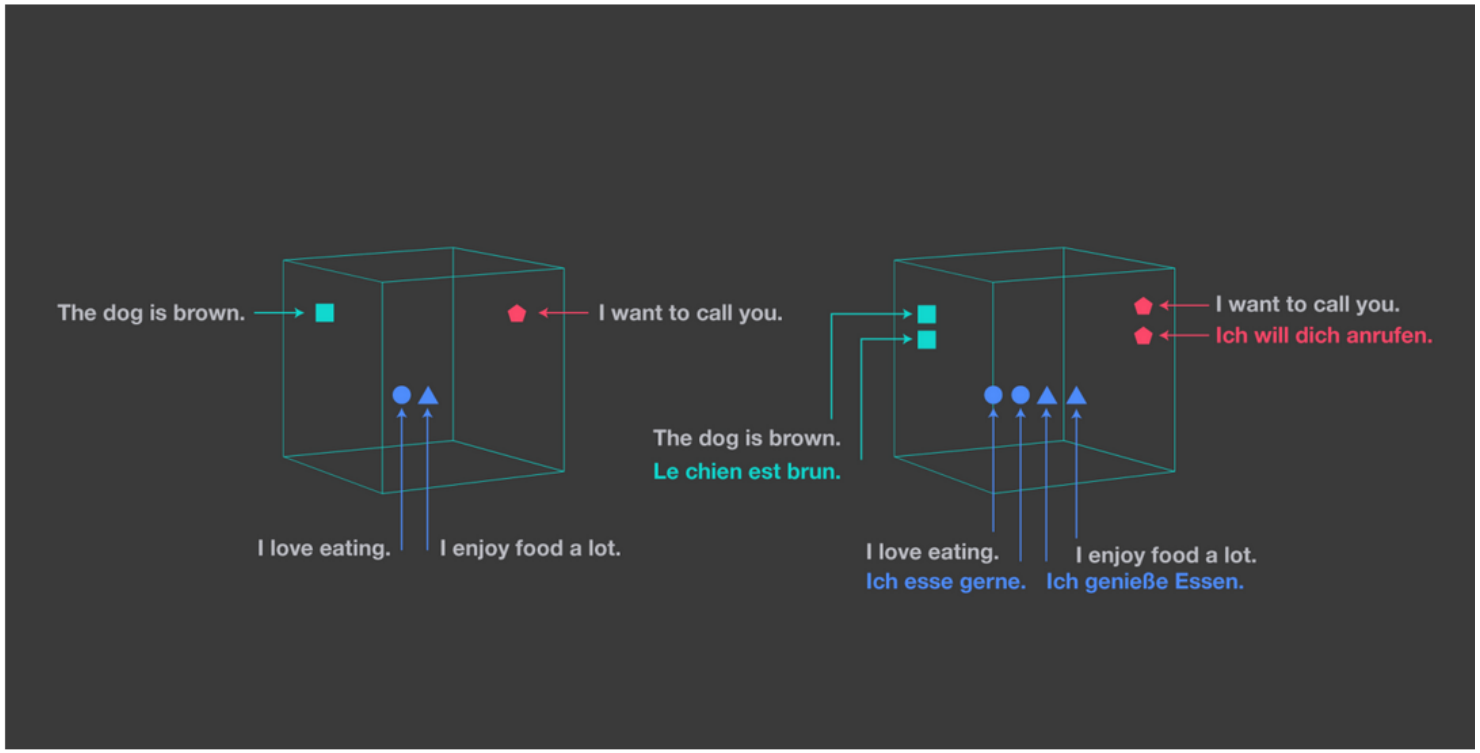
چالش های پیش پردازش

- نا متوزان بودن طول برخی از غزل ها
- وجود ایموجی حروف اضافه و بی معنی در عبارات های انگلیسی
- نرمالایز کردن عبارت های عربی نهج البلاغه دارای چالش های بسیاری بود
- استفاده از عبارت های منظم برای از بین بردن این مشکل ها

چالش اصلی مسئله

- ایجاد یک فضای Embedding چند زبانه

نمایشی از فضای embedding چند زبانه



مدل های زبانی مورد استفاده

- LaBSE
- LASER
- FastText (TF-IDF weighted)

الگوریتم استفاده شده برای پیدا کردن جفت های ورودی خروجی

- The K-nearest neighbors algorithm
- Cosine similarity for similarity metrics

LaBSE

- مدل document embedding pre trained هست که از کتابخونه hugging face sentence-transformers
- فضای مشترک تولید میکند
- Multi-lingual (۱۰۹ تا زبان)
- Vector-length:786

LASER

- Langiage agnostic
- Vector-length:1024
- Sentence Embedding

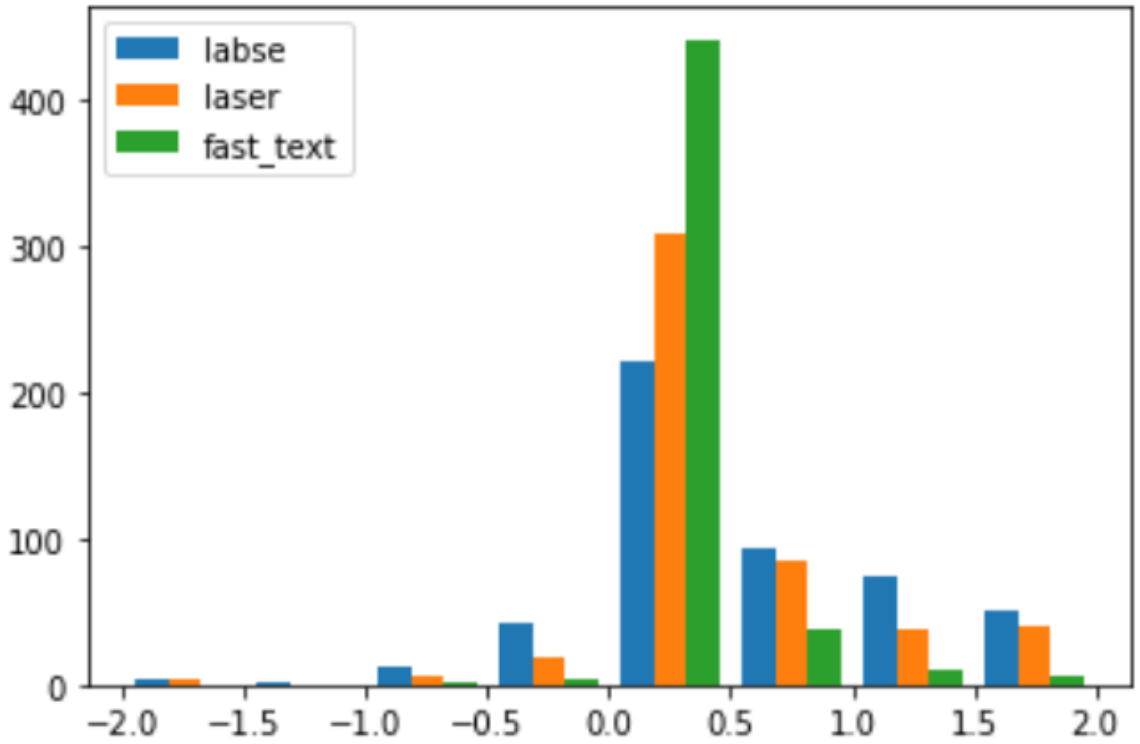
FastText (TF-IDF weighted)

- حساب کردن ضریب tf-idf هر یک از کلمات دیتاست ها و ورودی ها
- استفاده از مدل های زبانی عربی و فارسی و انگلیسی fasttext که روی ویکیپدیا این زبان ها آموزش داده شده اند
- Word Embedding
- Vector lengths:300
- محاسبه sentence vector با استفاده از جدول tf-idf و word-Embedding های مدل زبانی fasttext

Hand labeling the output

	input_id	output_id	input_text	output_text	chapter	tag1	tag2	-2	-1	0	1	2
0	0	5318	باید حقّوق دیگران را رعایت کنیم	... و قال ع من نصب نفسه للناس اماما فلیبدا بتعلیم	73	1	1	0	0	0	2	0
1	4	5564	نباید براساس ظاهر افراد را قضاوت کرد	...و قال ع للظالم من الرجال ثلاث علامات یظلم من ف	350	0	0	0	0	2	0	0
2	9	5563	باید قدردان دارایی‌های خودمان باشیم	و قال ع اشد الذنوب ما استهان به صاحبه	348	0	0	0	0	2	0	0
3	13	5563	باید بخشنده باشیم	و قال ع اشد الذنوب ما استهان به صاحبه	348	0	0	0	0	2	0	0
4	14	5414	باید حق‌جو وعدالت‌پیشه باشیم	و قال ع ما شککت فی الحق مذ اریته	184	2	1	0	0	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
1445	33	3406	نباید در خوردن زیاده‌روی کرد	You can never be overdressed or overeducated	NaN	0	0	0	0	2	0	0
1446	36	3172	باید اهل یاری کردن یکدیگر بود	...چون پرده براندازد عالم به سرانداز جایی که یقین	NaN	0	0	0	0	2	0	0
1447	3	5310	باید خاک وطن را دوست داشت	و قال ع فقد الاحبه غربه	65	0	1	0	0	1	1	0
1448	16	5592	نباید از دیگران درخواست های متعدد داشت	و قال ع من طلب شیئا ناله او بعضه	386	0	-1	0	1	1	0	0
1449	48	377	باید از منافق دوری کنیم	...ما نگوییم بد و میل به ناحق نکنیم جامه کس سیه و	غزل ۳۷۸	0	1	0	0	1	1	0

Evaluation



- LaBSE
- LASER
- FastText (TF-IDF weighted)

Fleiss Kappa score

```
fleiss_kappa(kapa)
```

```
0.2934708149418 92
```

- با استفاده از کتابخانه statsmodel

Other Evaluation metrics

- NDCG با استفاده از کتابخانه sklearn.metrics
- CG@1 با استفاده از کتابخانه sklearn.metrics
- MRR با استفاده از کتابخانه tensorflow و tensorflow_ranking

NDCG score for each model

```
LaBSE NDCG  Score is :
          tak abiat : 0.8
      english quotes : 0.8
        divane shams : 0.8
      qazaliat hafez : 0.78
```

```
LASER NDCG  Score is :
          tak abiat : 0.72
      english quotes : 0.78
        divane shams : 0.6
      qazaliat hafez : 0.5
```

```
Fast Text NDCG  Score is :
          tak abiat : 0.38
      english quotes : 0.16
        divane shams : 0.2
      qazaliat hafez : 0.26
```

CG@1 score for each model

LaBSE CG@1 Score is :		LASER CG@1 Score is :		Fast Text CG@1 Score is :	
tak abiat	0.4266666666666667	tak abiat	0.38	tak abiat	0.12666666666666668
english quotes	0.575	english quotes	0.545	english quotes	0.055
divane shams	0.515	divane shams	0.3	divane shams	0.07
qazaliat hafez	0.535	qazaliat hafez	0.2	qazaliat hafez	0.1
hekam nahjol balaghe	0.72	hekam nahjol balaghe	0.39	hekam nahjol balaghe	0.03

MRR score for each model

LaBSE mrr Score is :	LASER mrr Score is :	Fast Text mrr Score is :
tak abiat : 0.6	tak abiat : 0.21333334	tak abiat : 0.053333335
english quotes : 0.52	english quotes : 0.29	english quotes : 0.02
divane shams : 0.54	divane shams : 0.17	divane shams : 0.03
qazaliat hafez : 0.46	qazaliat hafez : 0.09	qazaliat hafez : 0.06
hekam nahjol balaghe : 0.4	hekam nahjol balaghe : 0.24	hekam nahjol balaghe : 0.0

Our Contribution

- یکی از چالش های اصلی پروژه ما نبود دیتاست supervised غنی سازی متن ورودی با شعر یا حکمت بود
- در همین راستا یکی contribution های اصلی ما در این پروژه جمع آوری ، مرتب سازی و تگ زدن داده های خروجی بر اساس داده های ورودی بود
- به عنوان خروجی پروژه ما دو دیتاست به فرمت فایل های تسک human value detection تولید کردیم
- این دیتا ست میتوانند داده های این تسک را بهبود دهند و هم برای ارزیابی و آموزش مدل های دیگر هم استفاده بشوند

Our Contribution

Whatever it is you're seeking won't come in the form you're expecting	به نفع	لزوما رسیدن به چیزی که آرزو داریم به نفع ما نیست
خاکم به سر ز غصه به سر خاک اگر کنم خاک وطن که رفت چه خاکی به سر کنم	به نفع	باید خاک وطن را دوست داشت
به همه حال شکر باید کرد که مبادا ز بد بتر گردد	به نفع	باید شکرگزار باشیم
هر کجا عشق آید و ساکن شود هر چه نا ممکن بود ممکن شود	به نفع	باید به قلب و احساسات توجه کرد
جزای حسن عمل بین که روزگار هنوز خراب میکند بارگاه کسری را	در مقابل	ثروت چیز خوبی است
گفتار بسیار نه از نغزیست ولوله طبل ز بی مغزیست	به نفع	کنترل نکردن زبان انسان را بی‌ارزش می‌کند.
جزای حسن عمل بین که روزگار هنوز خراب میکند بارگاه کسری را	به نفع	باید سعی کنیم نیکو عمل کنیم
به هوش باش دلی را بسهو نخراشی به ناخنی که توانی گره گشایی کرد	به نفع	باید به قلب و احساسات توجه کرد
و قال ع من وضع نفسه مواضع التهمه فلا یلومن من اساء به الظن	در مقابل	نباید براساس ظاهر افراد را قضاوت کرد
I decided it is better to scream. Silence is the real crime against humanity	به نفع	انسان واقعی در برابر ظلم سکوت نمی‌کند

Our Contribution

- ورودی : مردم را باید از کارهای بد باز داشت

عیبِ رندان مَکُن ای زاهدِ پاکیزه سرشت — که گناهِ دگران بر تو نخواهند نوشت

من اگر نیکم و گر بد تو برو خود را باش — هر کسی آن دِرَوَد عاقبتِ کار، که کِشت

همه کس طالبِ یارند چه هشیار و چه مست — همه جا خانهٔ عشق است چه مسجد چه کِنِشت

سرِ تسلیمِ من و خشتِ درِ میکده‌ها — مدعی گر نکند فهمِ سخن، گو سر و خشت

ناامیدم مکن از سابقهٔ لطفِ ازل — تو پسِ پرده چه دانی که که خوب است و که زشت

نه من از پردهٔ تقوا به درافتادم و بس — پدرم نیز بهشتِ ابد از دست بهشت

حافظا روزِ اجل گر به کف آری جامی — یک سر از کویِ خرابات بَرَندَت به بهشت

Reference

Feng, F., Yang, Y., Cer, D., Arivazhagan, N. and Wang, W., 2020. Language-agnostic bert sentence embedding. arXiv preprint arXiv:2007.01852.
Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, pp.135-146.
Xia, F. and Lewis, W., 2009, March. Applying NLP technologies to the collection and enrichment of language data on the web to aid linguistic research. In Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH–SHELT&R 2009) (pp. 51-59).
https://github.com/facebookresearch/LASER
https://bigdata-ir.com/
https://huggingface.co/datasets/Abirate/english_quotes
https://amitness.com/2020/08/information-retrieval-evaluation

Thank you