بسم الله الرحمن الرحيم

موضوع پروژه :

سیستم استخراج اصطلاحات تخصصی حوزه اقتصاد (مبتنی بر BERT)

ارائه دهندگان :

خانم نرگس علی حیدری

خانم فاطمه قیصری

آقای علیرضا فرزانه

استاد پروژه:

سرکارخانم دکتر پیشگو

موضوع درس:

یادگیری ماشین

مقدمه

در عصر انفجار اطلاعات، دسترسی سریع به مفاهیم کلیدی در یک حوزه تخصصی اهمیت فراوانی دارد ساخت واژه نامه‌های تخصصی به صورت دستی، فرآیندی زمان بر و پرهزینه خواهد بود . هدف این پروژه، خودکارسازی این فرآیند برای متون اقتصادی است تا بتوان با دریافت یک متن خام، لیستی از واژگان کلیدی آن را استخراج کرد. استخراج خودکار واژگان تخصصی (Automatic Glossary Extraction) یکی از چالش‌های کلیدی در پردازش زبان طبیعی است که هدف آن شناسایی اصطلاحات دامنه محور از متون غیرساختاریافته می‌باشد. در این پژوهش، یک (Pipeline) مبتنی بر مدل زبانی DistilBERT و الگوریتم‌های خوشه بندی برای استخراج واژگان حوزه اقتصاد طراحی شده است که برخلاف روش‌های سنتی مبتنی بر آمار (مانند TF-IDF)، روش glossex بر درک معنایی و برداری کلمات تمرکز دارد.

هدف پروژه

استخراج خودکار اصطلاحات تخصصی حوزه اقتصاد از یک متن (Corpus) است؛ به طوری که خروجی نهایی یک لیست رتبه بندی‌شده از "ترم‌های اقتصادی "باشد (شامل تک واژه و عبارت‌های چندواژه ای)، بدون این‌که دیتاست برچسب‌خورده‌ی بزرگ داشته باشیم.

Weakly Supervised (نظارت ضعیف)

این پروژه weakly supervised است زیرا به جای داشتن برچسب‌های کامل، از seed words (چند کلمه‌ی نمونه‌ی اقتصادی و چند کلمه‌ی عمومی) برای هدایت الگوریتم استفاده می‌کنیم و با سیگنال ضعیف (weak signal) دامنه را تشخیص می‌دهیم.

داده‌ها (Dataset)

برای استخراج واژگان، یک corpus در حوزه‌ی اقتصاد استفاده شده corpus شامل تنها چند جمله در حوزه‌ی اقتصاد (مانند مفاهیم عرضه و تقاضا، تورم، رشد اقتصادی و بازارها)، که می توان اندازه‌ی corpusرا افزایش داد در واقع افزایش اندازه‌ی corpus نقش مهمی در بهبود پوشش واژگان و پایداری مراحل بعدی pipeline خواهد داشت.

در این پروژه پیکره‌ی متنی اقتصاد در فایل زیر قرار دارد:

- data/raw/economics_sample.txt

این فایل ابتدا یک نمونه‌ی کوچک بود و سپس برای بهتر شدن خروجی، متن دامنه‌ای بزرگ تر به آن اضافه شد (افزایش حجم متن ← افزایش تنوع واژگانی ← کاندیدهای بهتر).

معماری کلی سیستم (Pipeline Overview)

پروژه به صورت یک خط لوله (Pipeline) طراحی شد که از خام‌ترین داده تا خروجی نهایی را تولید می‌کند:

1. **Preprocess**: پاکسازی، نرمال‌سازی، توکن‌سازی
2. **Baseline (TF-IDF)**: استخراج کلمات مهم بر اساس فراوانی/ تمایز
3. **Embedding**: ساخت بردار معنایی برای واژگان/ لمّاها با مدل زبانی
4. **Clustering**: خوشه‌بندی بردارها برای گروه بندی ترم‌های نزدیک معنایی
5. **Filtering (Seed-guided)**: انتخاب خوشه‌های اقتصادی با مقایسه شباهت به seedهای اقتصادی وseed های عمومی
6. **Phrase Extraction + Ranking** (فاز ۲): استخراج عبارت‌های چندکلمه‌ای n-gram و رتبه‌بندی هیبرید (TF-IDF + Embedding similarity)
7. **Evaluation**: ارزیابی درK های مختلف (200/100/50) نسبت به Gold Glossary

Phase1

در فاز ۱ یک اسکریپت EDA ساخته شد و خروجی نمودارها ذخیره شد مثل توزیع طول جمله، توزیع طول کلمات، Zipf، رشد واژگان، top frequent words و....
هدف EDA در گزارش:

- نشان دهد متن ورودی چقدر بزرگ است،
- تنوع واژگان چقدر است،

- آیا متن واقعاً دامنه‌ای است یا نه،
- آیا داده برای استخراج واژگان کافی هست یا باید غنی‌تر شود.

خروجی‌ها و مستندسازی فاز ۱

در فاز ۱ موارد زیر در پروژه قرار دارد:

- گزارش فاز ۱ docs/phase-1-report.md
- ثبت نتایج baseline و برنامه ی فاز ۲(experiment plan)
- ذخیره نمودارهای EDA در مسیر نتایج

فاز 2 :(Phase-2) روش پیشنهادی + Weakly Supervised بهبودها + ارزیابی

فاز 2 روی "روش اصلی" و "بهبود و مقایسه" تمرکز دارد.

ایده‌ی weak supervision در این پروژه از طریق لیستseed ها پیاده‌سازی شد:

data/seeds/economics.txt نمونه‌های اقتصادی مثل (economics, market) .

data/seeds/general.txt نمونه‌های عمومی مثل(education, school……)

مراحل فاز 2 :

preprocess

در مرحله اول

- متن به توکن‌ها شکسته شد
- نرمال‌سازی اولیه انجام شد
- توکن‌های نامعتبر حذف شدند

پس از نرمال‌سازی و حذف توکن‌های نامعتبر خروجی این قسمت وارد بخش embedding می شود.

پس از پیش‌پردازش، مشخصات داده به صورت زیر بود:

- تعداد توکن‌های نهایی: 8981 token

خروجی : data\processed\preprocess.json

saliency scoring

در این مرحله میزان تخصصی بودن هر کلمه سنجیده می شود . برای این کار از مقیاس zipfاستفاده شده است تا فرکانس کلمه در متن پروژه با فرکانس آن در زبان انگلیسی عمومی مقایسه شود کلماتی که در متن ما تکرار بالا و در زبان عمومی تکرار پایینی دارند امتیاز بیشتری می گیرند

```
score = math.log(domain_prob / general_prob + 1e-9)
```

خروجی : demo\output\top_saliency_terms.csv

Embedding

در این قسمت از تکنولوژی ترنسفورمراستفاده شده هر کلمه ورودی رو به یک (vectot) یک لیست طولانی از اعداد در فضای ریاضی تبدیل می کند در مثال آورده شده هر 8918 توکن حالا در فضای ریاضی یک موقعیت و معنایی دارند ما 2787 بردار embeddingداریم. این کاهش امبدینگ ها به دلیل حذف کلمات تکراری - حذف stopword- ریشه یابی یا lemmatizatioاست. استفاده از مدل BERT در embeddings.py برای تبدیل کلمات به بردارهای ۷۶۸ بعدی است این مرحله باعث می‌شود کلماتی مثل ”Inflation“ ”economic“ در فضای ریاضی به هم نزدیک شوند.برخلاف مدل های word2vec این بردارها حاوی اطلاعات contex هستند که دقت خوشه بندی رو به شدت افزایش می دهد .

تعداد امبدینگ های نهایی : 2784 embedding

خروجی : data\processed\lemma_embeddings.json

Clustering

توکن‌ها بر اساس شباهت برداری خوشه بندی شدند تا واژگان هم معنی در کنار یکدیگر قرار گیرند.یعنی کلماتی که بردارهای آنها نزدیک به هم بوده در یک خوشه قرار گرفتند مثل , "inflation " "economic"" در نهایت در این پروژه 696 خوشه تشکیل شد که هر کدام نماینده یک مفهوم بالقوه بودند در این مرحله از Agglomerative که یک خوشه بندی سلسله مراتبی می باشد استفاده شده است.

تعداد خوشه های نهایی : 696 cluster

خروجی: data\processed \clusters.json

Filtering

برای تشخیص خوشه‌های مرتبط با اقتصاد، از یک روش مبتنی بر seed words استفاده شد:

در این قسمت از دو seed استفاده کردیم

- **Economic seeds**: واژگان مرتبط با اقتصاد
- **General seeds**: واژگان عمومی و غیرتخصصی

این دقیقاً همان جایی است که پروژه weakly supervised می‌شود: ما برچسب دقیق نداریم، ولی با یک سیگنال کم‌هزینه (seed list) جهت‌دهی می‌کنیم.

در filtering با مقایسه شباهت خوشه‌ها به Seedهای اقتصادی، واژگان نهایی استخراج می‌شوند.

اگر میانگین شباهت اقتصادی > عمومی باشد، خوشه "اقتصادی" در نظر گرفته می‌شود

فرمول محاسبه شده در پروژه cosine similarity می باشد

$$Cosine = a.b / |a|.|b|$$

خروجی: data\processed\final_terms.json

phrases

برای نزدیک شدن به glossary واقعی که معمولاً multi-word term است در واقع ما همیشه با تک واژه ها رو به رو نیستیم می خواهیم اصطلاحات تخصصی رو استخراج کنیم در این صورت به وسیله ngramها که ما در این پروژه bigram , trigram داریم عبارات رو استخراج می کنیم.

- استخراج کاندیدهای عبارتی با n-gram + TF-IDF انجام شد

خروجی : data\processed\phrase_candidates.json

ranked_phrases

میانگین بردار embeddingهای مرحله ghrases را بدست آورده و میانگین شباهت آن ها با seed economics و seed general از طریق شباهت cosine انجام شده و scoreخروجی بررسی می شود اگر میانگین شباهت اقتصادی > عمومی باشد، خوشه "اقتصادی" در نظر گرفته می‌شود

خروجی : data\processed\rank_phrases.json

Hybrid Ranking

ترکیب TF-IDF و Embedding

برای بهتر شدن دقت نسبت به:TF-IDF:

- یک امتیاز هیبرید تعریف شد که هم "اهمیت آماری (TF-IDF) "و هم "شباهت معنایی به seedهای اقتصادی" را لحاظ می‌کرد.
- این باعث شد خروجی‌ها از حالت صرفاً "کلمات پرتکرار عمومی" فاصله بگیرند و اقتصادی‌تر شوند.
- این ترکیب طبق فرمول زیر محاسبه میشود

Hybrid=alpha*tfidf_n +(1-alpha)*emb_n

الفا =1-0

آلفا : 0.8 به دلیل قابلیت اطمینان بالاتر

خروجی: data\processde\ranked_hybrid.json

Baseline

به عنوان معیار مقایسه baseline ، از روش TF-IDF برای استخراج واژگان پرتکرار استفاده شد. این روش بدون درک معنایی embeding، صرفاً بر اساس فراوانی و تکرار واژگان عمل می‌کند.

خروجی: demo\outputs\tfidf_baseline_top_terms.csv

Evaluation

در فاز ۲ اسکریپت ارزیابی طوری به روزرسانی شد که در چند K گزارش بدهد:

K=200	K=100	K=50	Glossex model
0.045	0.040	0.060	precision
0.090	0.040	0.030	recall
0.060	0.040	0.040	F1

K=200	K=100	K=50	Tfidf model
0.010	0.020	0.040	precision
0.020	0.020	0.020	recall
0.013	0.020	0.027	F1

تحلیل نتایج:

- درK های بزرگـتر، Recall روش پیشنهادی افزایش پیدا می‌کند (یعنی اصطلاحات مرجع بیشتری پوشش داده می‌شود).
- روش پیشنهادی نسبت به baseline در اکثرK ها بهتر است خصوصاً در.Recall
- این با هدف glossary extraction هم سازگار است: در بسیاری از سناریوها Recall مهم‌تر است چون حذف موارد بد ساده تر از پیدا کردن اصطلاحات جا افتاده است.

ساختار نهایی (Project Structure)

- data داده خام
- src کد اصلی پروژه pipeline ها، baseline،
- scripts اسکریپت‌های اجرایی تولید خروجی
- evaluation ارزیابی دو مدل پایه و مدل فاز 2 پروژه
- docs گزارش‌های فاز ۱ و فاز ۲
- demo/outputs خروجی‌های نمایشی
- notebooks تحلیل‌ها و EDA تعاملی

جمع‌بندی نهایی

در این پروژه یک سیستم استخراج واژگان اقتصاد ساخته شد که:

- از متن دامنه‌ای اقتصاد به عنوان ورودی استفاده می‌کند،
- با baseline کلاسیک TF-IDF شروع می‌کند،
- سپس با embedding + seed- filtering و hybrid ranking به سمت weak supervisionمی‌رود،
- و با معیارهای Precision / Recall در چند K ارزیابی می‌شود.

نتیجه‌ی نهایی نشان داد روش پیشنهادی در سناریوهای مختلف به‌خصوص از نظر Recall نسبت به baselineبهبود دارد و به هدف glossary extraction نزدیک تر است.

References

- Wikipedia – Economics برای تعریف و توضیح کلی حوزه (Encyclopedia Britannica) .
- Encyclopaedia Britannica – *The unintended effects of markets* برای محتوای دامنه‌ای و غنی‌سازی متن اقتصاد(Wikipedia) .
- Velardi et al., IEEE Intelligent Systems (2008)درباره اهمیت و معماری کلی Glossary Extractionو چرایی نیاز به اتوماسیون.