CSE4078 Introduction to Natural Language Processing

NLP Project Delivery #1
NLP Task and Dataset Report

Group Number 10

150119880 Abdullah Enes Dizer

150119919 Ayberk Türksoy

150119805 Mert Efe Karaköse

150119751 Musa Meriç

150119715 Yağmur Koçoğlu

NLP (Natural Language Processing) is a sub-branch of artificial intelligence used for computers to understand and process human language. It usually works on textual data and is used to make sense of, classify or infer this data. NLP has many application areas and one of them is classification/categorization.

Classification/categorization is a sub-branch of NLP and generally involves the process of assigning text data to specific categories or classes. This process aims to analyze the content of a document or text and place it into the most appropriate category or class. This process is supported by machine learning and deep learning techniques.

The categories we set for NLP classification/categorization:

1-Dizi/Film: Series and film reviews can be assigned to this category, taking into account text data, content analysis and emotional tone. For example, a film review belongs to the category "Dizi/Film".

2-Ekonomi: Financial reports, news articles or economic analyses can be assigned to this category. For example, an economics article belongs to the category "Ekonomi".

3-Dünya: International news, political analyses or geographical topics can be assigned to this category. For example, a world developments report belongs to the category "Dünya".

4-Sağlık: Health news, medical research or health information texts can be assigned to this category. For example, a health information article belongs to the category "Sağlık".

5-Teknoloji: Technology news, product reviews or computer science research can be assigned to this category. For example, a new technology product white paper belongs in the "Teknoloji" category.

6-Şikayet/ Müşteri Geri Bildirim: Customer complaints, service evaluations or consumer feedback can be assigned to this category. For example, a customer complaint belongs to the category "Şikayet/ Müşteri Geri Bildirim".

7-Magazin: Magazine Contains popular culture news and celebrity interviews. For example, a celebrity interview belongs to the "Magazine" category.

8-Kültür: Culture Includes art events and cultural topics. For example, a review of an art exhibition belongs to the category "Kültür".

9-Spor: Sports news, match reports or sport events can be assigned to this category. For example, a football match report belongs to the category "Spor".

10-Ürün Yorumu: Product reviews, user comments or product comparisons can be assigned to this category. For example, an electronic device review belongs to the category "Ürün Yorumu". The categories mentioned in these examples are examples of text data that are typically used for NLP classification/categorisation and that are to be classified. Analysing the features specific to these categories is important in building the correct classification model

	row number	dataset name	url	source	description	#of instances
1	14.5k	Ttc4900	link	Hugging Face	A text classification dataset with 7 different news category.	4.9k
2	211k	42000 Turkish News Texts in 13 Classes	link	Kaggle	The dataset contains 42000 Turkish news texts in 13 classes. There is a news folder that contains 13 classes, each folder classified with the topic name.	42k
3	41.6k	TRT-Haber-World-for-News-Pages-Data	link	Kaggle	TRT Haber news	8328
4	3.730	TRT_Haber_Ekonomi	link	Kaggle	TRT Haber Economy	746
5	20.5k	Turkish News Dataset	link	Kaggle	Dataset consists of 7 news type labels. These labels are economy, politics, life, technology, magazine, health, sport.	4115
6	8.5k	TRT-Haber-Kültür-Sanat	link	Kaggle	TRT Haber Culture	1700
7	176.5k	beyazperde-all-movie-reviews	link	Hugging Face	Beyazperde Movie Reviews Offers Turkish sentiment analysis datasets that is scraped from popular movie reviews website Beyazperde.com.	35.3k
8	220k	beyazperde-top-300-movie-reviews	link	Hugging Face	Top 300 Movies include audience reviews about best 300 movies of all the time.	44k

	row number	dataset name	url	source	description	#of instances
9	1.175m	turkish_product_reviews	link	Hugging Face	This Turkish Product Reviews Dataset contains 235.165 product reviews collected online.	235k
10	16.9k	turkishReviews-ds-mini	link	Hugging Face	This dataset contains data on the category of customer complaint feedback in Turkish.	3.38k

EXAMPLES in ALPACA FORMAT

```
{
      "instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin kategorilerinden hangisine ait olduğunu söyle.",
      "input": "oyunculuk güzel ama sadece oyunculuk... konusunu hiç beğenmedim...",
      "output": "Bu cümle 'Dizi-Film' kategorisine aittir."
  },

{
      "instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin kategorilerinden hangisine ait olduğunu söyle.",
      "input": " hepatit c tarih olabilir ! uzmanlardan hepatit müjdesi ankara_üniversitesi tıp_fakültesi iç_hastalıkları ve gastroenteroloji bilim dalı öğretim üyesi_prof dr hasan özkan bu hastalığın tedavisiyle ilgili olan ve onay alan iki ilaçta yüzde 95 in üzerinde başarılı sonuçların alındığını söyledi 9 ulusal hepato gastroenteroloji kongresi nde konuşan özkan şunları anlattı yakın gelecekte hepatit c tarihe karışacak günümüzde hastaların yüzde 40 ı şifa buluyor ancak şimdi onay alan iki yeni ilaçla tedavide çok daha başarılı sonuçlar elde edebileceğiz özkan karaciğer konusunda sadece enginar ve devedikeninin yararlı olduğunun belirlendiğini de söyledi",
      "output": "Bu cümle 'Sağlık' kategorisine aittir."
  }
```

How to Convert to Alpaca format?

- **change_instruction.py :** modifies the instructions in the code
- **csv-to-json.py :** converts csv file to json file
- **json-to-json.py :** converts from json format to more regular json format
- **jsonFormatEdit.py :** converts json format to the format we want
- **jsonFormatFiltering.py :** extracts unwanted places in the json file
- **txtToJson.py :** converts txt file to json file