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.
11-Kitap: News about books, book summaries or book reviews can be assigned to this category. For example, a review of an author's new book could belong to the “Kitap” category.

	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

11	6.7k	TurkishBookDataSet	link	Kaggle	This dataset is a comprehensive compilation of Turkish books obtained through web scraping from the internet.	1.34k

Examples in Alpaca Format

Dataset1

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "oyunculuk güzel ama sadece oyunculuk... konusunu hiç beğenmedim...",

"output": "Bu cümle 'Dizi-Film' kategorisine aittir."

},

Dataset2

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "Büyükşehir Belediyesi fuaye alanında, Sıfır Atık Projesi kapsamında düzenlenen sergide, eski mobilyalar üzerine kursiyerler tarafından el boyama, eskitme, altın varak, rölyef pasta gibi teknikler uygulanarak yapılan 12 takım ve 70 parça ürün yer aldı. Serginin açılışına Kayseri Büyükşehir Belediye Başkanı Memduh Büyükkılıç, Özvatan Kaymakamı Eray Gürsoy, AK Parti İl Kadın Kolları Başkanı Emine Timuçin, Erciyes Üniversitesi Güzel Sanatlar Fakültesi Dekanı Prof. Dr. Aygül Aykut ve KAYMEK AŞ Yönetim Kurulu Başkanı Serdar Öztürk katıldı.\"",

"output": "Bu cümle 'Kültür' kategorisine aittir."

},

Dataset3

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "Doç. Dr. M. Ali Kaplan, meme kanserinin tedavisinde uygulanan akıllı ilaçların, tedavide başarı sağladığını söyledi",

"output": "Bu cümle 'Sağlık' kategorisine aittir."

},

Dataset4

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "Dropbox kendini dörde katladı Popüler bulut servisi Dropbox, 100 milyon aboneye ulaştığını duyurdu Geçen yıla göre abona sayısını neredeyse dörde katlayan Dropbox, özellikle Android tabanlı telefonlara sunulmasıyla birlikte kullanıcı kitlesini hızla genişletmeye başladı 2 GB'a kadar ücretsiz olarak kullanılabilen Dropbox'ta daha fazla kapasite kullanmak isteyenler ücretli aboneliğe geçiş yapıyor Özellikle Google'ın rakip servisi Google Drive'ın yayınlanmasıyla rekabet gittikçe kızışırken, Dropbox aslında sadece Google Drive değil, Microsoft'un SkyDrive, Apple'ın iCloud bulut servislerine karşı da mücadele veriyor.",

"output": "Bu cümle 'Teknoloji' kategorisine aittir."

},

Dataset5

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "Borsada işlem gören 169 şirket 3'üncü çeyrek bilançosunu açıkladı.",

"output": "Bu cümle 'Ekonomi' kategorisine aittir."

},

Dataset6

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": " mursi den muhalefete diyalog çağrısı mısır_cumhurbaşkanı_muhammed_mursi yetkilerini genişleten kararnameye yönelik tepkiler üzerine diyalog adımı attı mısır_cumhurbaşkanı_muhammed_mursi kararlarının yargıdan muaf olması yönünde çıkardığı kararnameye gelen tepkiler üzerine diyalog çağrısı yaptı mursi yetkilerinin artırılmasının geçici olduğunu belirtti ve ortak bir zemin bulunması için diyalog çağrısında bulundu bu çağrının ardından mursi nin yarın ülkenin en yüksek yargı organı olan yüksek yargı konseyi üyeleriyle görüşeceği açıklandı bu arada mursi nin kararını protestolar devam ediyor kahire deki tahrir_meydanı na yakın bölgelerde göstericilerle polis arasında çatışmaların yer yer devam ettiği belirtiliyor gösterilerin iskenderiye ismailiye ve port_said kentlerinde de devam ettiği gelen bilgiler arasında yaklaşık 300 kişinin gözaltına alındığı belirtilirken tahrir_meydanı ndaki oturma eylemi de sürüyor ntvmsnbc",

"output": "Bu cümle 'Dünya' kategorisine aittir."

},

Dataset7

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "Halkbank İhtiyaç Kredim Halen Onaylanmadı. Ben bundan 5 gün önce arkadaşım ile aynı anda kredi başvurusu yaptım, onunki ertesi gün onay geldi ama bana gelmedi. Ben anlamadım yani bu işi adamına göre muamele mi yapıyorsunuz anlamıyorum. Şikayette bulunmuştum zaten o bile görülmemiş onu sildim şimdi tekrar bulunuyorum. Allah aşkına bir yardımc",

"output": "Bu cümle 'Şikayet-Müşteri Geri Bildirim' kategorisine aittir."

},

Dataset8

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "Rol aldığı dizi geçtiğimiz hafta final yapan oyuncu Hazar Ergüçlü, önceki gün Atatürk Havalimanı'nda görüntülendi.",

"output": "Bu cümle 'Magazin' kategorisine aittir."

},

Dataset9

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "TVF'den Fenerbahçe Kulübü'ne para cezası Türkiye Voleybol Federasyonu (TVF) Ceza Kurulu, Fenerbahçe Grundig'in Galatasaray'ı konuk ettiği Erkekler Voleybol 1 Lig 3 hafta maçında, taraftarların neden olduğu olaylar nedeniyle sarı-lacivertli kulübe para cezası verdi.TVF tarafından yazılı olarak Fenerbahçe Kulübü'ne gönderilen açıklamada, 28 Ekim Pazar günü Burhan Felek Voleybol Salonu'nda oynanan derbi maçta, taraftarların neden olduğu olaylar nedeniyle, sarı-lacivertli kulübe 20 bin lira para cezasının verildiği bildirildi.",

"output": "Bu cümle 'Spor' kategorisine aittir."

},

Dataset10

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "bu fiyata çok iyi",

"output": "Bu cümle 'Ürün Yorumu' kategorisine aittir."

},

Dataset11

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "parasını karşılığını hakeden ürünlerden biri",

"output": "Bu cümle 'Ürün Yorumu' kategorisine aittir."

},

Dataset12

{

"instruction": "Aşağıdaki cümlenin Dünya, Ekonomi, Kültür, Sağlık, Spor, Teknoloji, Ürün Yorumu, Dizi-Film, Şikayet-Müşteri Geri Bildirim, Magazin, Kitap kategorilerinden hangisine ait olduğunu söyle.",

"input": "İyi Geceler Tatlı Prens, bir baba-oğul arasındaki sevginin, yanlış anlamaların ve söylenmemiş sözlerin hikâyesi. Ölmüş babasının ardından oğulun gırtlağından yükselen hüzünlü bir çığlık. Yaşam, zaman, sevgi ve ölüm üzerine sarsıcı ama gene de alaycı bir incelikle aktarılan duygular. Babasını kaybetmiş, ona hayattayken bir kez olsun \"Seni çok seviyorum\" diyememiş olan oğulun, duyduğu acı ve özlemi giderebilirmişçesine geçmişi sözcüklerle yeniden kurmaya, babasına söyleyemediklerini, geç de olsa bu yolla dile getirmeye çalışması...Pierre Charras, örnek, ideal bir evlat tipi yaratmanın, baba-oğul ilişkilerini evrensel boyutta irdelemenin peşinde değil. Bir baba ve oğul arasındaki, sevgi ve suskunluklarla örülü, son derece kişisel denebilecek ilişkiyi, okurun yüreğine işleyen bir şiirsellikle aktarıyor. Sevginizi dile getirmeyi, yaşamı ertelemeyin diyor bize, yoksa geç kalabilirsiniz...Sayfa Sayısı: 104Baskı Yılı: 2010Dili: TürkçeYayınevi: Can Yayınları",

"output": "Bu cümle 'Kitap' 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
- **csv2t_to_json:** converts csv file to json file
- **json2_to_json:** converts from json format to more regular json format
- **calculate_character_count:** calculates the number, mean and standard deviation of characters in input, output and instruction

STATISTICS

number of total row = 1.829.185

number of instruction = 365.837

number of input = 365.837

number of output = 365.837

Input

total number of input characters = 145.898.111

avg length of input characters = 398,81

std dev = 775,74

Output

total number of output characters = 15.356.185

avg length of output characters = 41,98

std dev = 2,71

Instruction

total number of instruction characters = 67.314.008

avg length of instruction characters = 184

std dev = 0