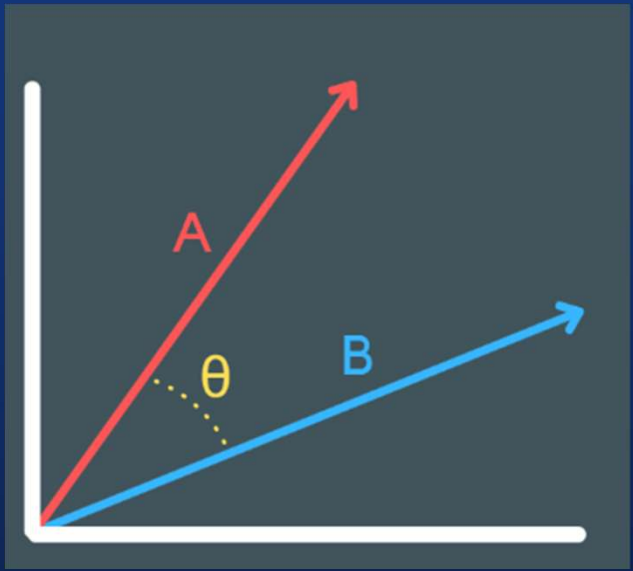
CLEF Simpletext Tasks 1-3

Tomislav Mikulandrić, University of Split, Faculty of Science
Rowan Mann, Christian-Albrechts-Universität zu Kiel (CAU)

SimpleText

ElasticSearch to retrieve documents

LLMs to rewrite, explain and identifly difficult terms



Task 1: “What is in (or out)?”

- Using the ElasticSearch API to retrieve abstracts
- Take the first 5 entries from task1_queries and call the ES API for each of them
- TF-IDF vectorization
- rel_score based on cosine similarity of query and abstract

```
# Function to query the ElasticSearch
def query_elasticsearch(query, size=100):
    response = requests.get(f"{ES_URL}?q={query}&size={size}", auth=('inex', 'qatc2011'))
    if response.status_code == 200:
        return response.json()['hits']['hits']
    else:
        print("Failed to fetch data:", response.status_code)
        return []
```

```python
def flesch_kincaid_grade_level(text):
    # Constants for the formula
    ASL = average_sentence_length(text)
    ASW = average_syllables_per_word(text)

    # Calculating the score
    score = 0.39 * ASL + 11.8 * ASW - 15.59

    # Normalize score to range from 0 to 1
    normalized_score = normalize(score, min_score=0, max_score=25)  # 20+ is academic level texts

    return normalized_score

def normalize(value, min_score, max_score):
    # Normalize value to range from 0 to 1
    return (value - min_score) / (max_score - min_score) if max_score != min_score else 0.5

def average_sentence_length(text):
    sentences = text.split('.')
    num_sentences = len(sentences)
    words = text.split()
    num_words = len(words)

    return num_words / num_sentences

def average_syllables_per_word(text):
    words = text.split()
    total_syllables = 0
    for word in words:
        total_syllables += count_syllables(word)

    return total_syllables / len(words)
```

Task 1: “What is in (or out)?”

- comb_score based on FKGL – take into account complexity
- Abstracts are usually on a College graduate level
- Need to simplify them for general public

Task 2: “What is unclear?”

- Using LLMs to analyze texts
- Using langchain replicate (ran out of tokens…)
- Switched to LLAMA2_13B_CHAT
- One shot prompting for difficulty
- Few shot prompting for explanations
- Create a list of difficult terms using LLAMA, split them and then ask the LLAMA model to provide explanations

```
prompt_difficulty="""
    You are a robot that rates the difficulty of different terms.
    You provide ONE LEVEL o difficulty for scientific terms.
    You need to consider two words as one term.
    Provide ONE rating for the understablity difficulty of term provided.
    There are 3 levels. You need to use: e for easy, m for medium and d for difficult.
    Give the rating inside of curly braces like this {e}
    You can reply with ONLY one word.
    Example source: autonomous vehicles
    Example answer: {'m'}
    Now here is my sentence:
"""
```

```
prompt_explanation="""
    You are a robot that explains difficult scientific terms.
    DO NOT add intro like "Sure, I'd be happy to help!"
    Use only once sentance and wrap the sentance in curly braces.
    Don’t justify your answers. Don’t give information not mentioned in the CONTEXT INFORMATION.
    Example source: wireless network environment
    Example answer: {'a system in which devices makes use of Radio Frequency connections between nodes in the network a system in which devices are co
    Example source: Bluetooth wireless technology
    Example answer: {'short-range wireless communication technology that allows devices to connect and exchange data. It facilitates data exchange bet
    Example source: application
    Example answer: {'software program or tool designed to perform specific tasks or functions on electronic devices. It can range from productivity t
    Example source: PDA
    Example answer: {'PDA is the acronym for personal digital assistant, which is a handheld electronic device designed for personal organization, com
    Example source: pilot study
    Example answer: {'a preliminary research investigation conducted on a small scale to assess the feasibility, and potential challenges of a larger
    Now here is my ONE sentence explanation:
"""

test.loc[test['difficulty'] == 'd', 'explanation'] = test.loc[test['difficulty'] == 'd', 'term'].apply(lambda x: completion(prompt_explanation + x))
test
```

Task 2: “What is unclear?”

- Wiki library for definitions (not working well considering terms marked as difficult are not available)
- Using regex to process the output of LLAMA model to make text seem „human like”

```python
import re

def remove_redundant_text(text):
    # Define patterns to search for
    patterns = [
        r'^Hey there!',
        r'^Sure!',
        r'^As a scientific journalist,',
        r'I\'m here to break down a complex study into simple terms for you\.',
        r'Here\'s a simplified version of the text',
        r'Let me break it down for you:',
        r'I\'m here to break down a complex study into simple terms for you\.',
        r'I\'m here to break down complex scientific concepts into simple, easy-to-understand language.',
        r'I\'m here to break down a complex topic into simpler terms for you. So, let\'s talk about',
        r'Here is my one sentence explanation of'

    ]
    # Compile regular expressions
    regex_patterns = [re.compile(pattern) for pattern in patterns]

    # Remove patterns from text
    for pattern in regex_patterns:
        text = re.sub(pattern, '', text).strip()

    return text
```

```python
[ ]  test = test.head(5)
     test['all_terms'] = test['source_snt'].apply(lambda x: extract_terms_from_string(completion(prompt_terms + x)))
     test = test.dropna(subset=['all_terms'])
     test
```

```python
[ ]  test['term']=test['all_terms'].str.split(";")
     test=test.explode('term').reset_index(drop=True)
     test.drop_duplicates(inplace=True,subset=['snt_id','term'])
     test
```

Rewrite this! Rewriting scientific text

```python
from ctransformers import AutoModelForCausalLM

# Load the model with a large context window
model = AutoModelForCausalLM.from_pretrained(
    "TheBloke/Llama-2-7B-32K-Instruct-GGUF",
    model_file="llama-2-7b-32k-instruct.Q4_K_M.gguf",
    model_type="llama",
    gpu_layers=50  # Adjust the number of GPU layers as needed
)
```

```python
def simplify(snt):
  c=model.create_chat_completion(
      messages = [
          {"role": "system", "content": "You are a scientific journalist who popularizes scientific results."},
          {
              "role": "user",
              "content": "Simplify the following text:\n"+snt
          }
      ]
  )
  return c['choices'][0]['message']['content'].strip()
```

- llama-2-7b-chat, increased the context window so that abstracts can fit
- Possibility to also use other LLAMA models with larger context, 32k tokens
- Could have also used context splitting
- Loaded last 25 sentances from the train set

Rewrite this! Rewriting scientific text

```python
def remove_redundant_text(text):
    # Define patterns to search for
    patterns = [
        r'^Hey there!',
        r'^Sure!',
        r'^As a scientific journalist,',
        r'I\'m here to break down a complex study into simple terms for you\.',
```

- Prompting the model to simplify the sentence
- Removing „fluff” again
- Formatting to json
- Same approach for abstracts

	query_id	query_text	doc_id	abs_id	source_abs	simplified_abs
155	M1	Alcohol interfer with recovery and adaptation ...	3	M1_3	Muscle contraction and the intake of leucine-r...	Muscles contract and consume leucine-rich prot...
156	M1	Alcohol interfer with recovery and adaptation ...	4	M1_4	The ingestion of ~20–25 g of high quality prot...	Eating about 20-25 grams of high-quality prote...
157	M1	Alcohol interfer with recovery and adaptation ...	6	M1_6	The cultural environment surrounding some spor...	Many athletes in team sports consume excessive...
158	M1	Alcohol interfer with recovery and adaptation ...	7	M1_7	The outcomes of binge drinking after exercise ...	Binge drinking after exercising can have two m...

	query_id	query_text	doc_id	snt_id	source_snt	simplified_snt
933	G01.1	Digital assistant	1533716782	G01.1_1533716782_2	We are interested in studying the effect of us...	:\nResearchers want to know how people's behav...
934	G01.1	Digital assistant	1533716782	G01.1_1533716782_3	We discuss three common transformation approac...	the three common approaches for displaying web...
935	G01.1	Digital assistant	1533716782	G01.1_1533716782_4	We introduce a new Overview method, called the...	Introducing "Gateway": A New Way to Browse Sci...
936	G01.1	Digital assistant	1533716782	G01.1_1533716782_5	The users in an initial study prefer using the...	In a recent study, participants preferred usin...
937	G01.1	Digital assistant	1534162055	G01.1_1534162055_1	The limitations and constraints of mobile syst...	In order to create effective software, it's im...

Questions?