GreekQA Dataset

Bonus Εργασία

Natural Language Processing (NLP)

- How to program machines to **process** and **analyze human language**
- Plethora of tasks:

  Language Modeling, Question Answering, Chatbots, Machine Translation, Data Augmentation, Text Classification, Text Generation, Sentiment Analysis, Name Entity Recognition, Word Embeddings, Text Summarization, Information Retrieval, Dialogue, Semantic Parsing etc.

More on Artificial Intelligence II class!

Language Modeling (LM)

- A task of **Natural Language Processing**
- Language Modeling is the task of **predicting what word comes next**
- A system that does this is called a **Language Model (LM)**
- You can also think of a Language Model as a system that assigns a **probability to a piece of text**

e.g. The students opened their _____ (*Books? Laptops? Minds? Doors?)*

More on Artificial Intelligence II class!

Question Answering (QA)

- A task of **Natural Language Processing**
- Question Answering is the task of **reading and comprehending a given text passage,** and then **answer questions based on it.**
- We often use **pre-trained Large Language Models (LLM)** for this task!
- These **LLMs** require further training *(fine-tuning)* on **QA datasets**.

More on Artificial Intelligence II class!

Stanford Question Answering Dataset (SQuAD)

- **Selected** high-quality **articles** from English Wikipedia
- **Extracted paragraphs (passages)** from these articles
- **Collected questions and answers** regarding these paragraphs
- **Collected additional answers** to the same collected questions

Why additional answers?

To evaluate *human performance* (and compare with *machine performance*!)

Every answer must be a span of text in the paragraph itself!
→ **Extractive Question Answering**

Stanford Question Answering Dataset (SQuAD)

Question: Which team won Super Bowl 50?

Paragraph (Answer):

> Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

Answer must be a span of text in the paragraph!

→ **Extractive QA dataset**

Extractive QA datasets

Dataset	Language	Paragraphs Source	Num. of QAs
SQuAD1.1	English	English Wikipedia	100k+
SQuAD2.0	English	English Wikipedia	150k+
FQuAD1.1	French	French Wikipedia	60k+
FQuAD2.0	French	French Wikipedia	79k+
KorQuAD1.0	Korean	Korean Wikipedia	70k+
KorQuAD2.1	Korean	Korean Wikipedia	102k+

Greek Question Answering Dataset (GreekQA)

- **Selected** high-quality **articles** from Greek Wikipedia
- **Extracted paragraphs (passages)** from these articles
- **Collected questions and answers** regarding these paragraphs

→ *Από bonus εργασίες σε άλλα μαθήματα*

- What else do we need?

Additional Answers!

Additional Answers

For **each paragraph**, we need **an additional answer to each of the given questions.**

- **Each answer:**
  - must be the **smallest possible text** in the paragraph that answers the question

Note: Each paragraph is assigned to two different students in order to collect two additional answers to each question

Examples of Answers

Each answer must be the **smallest possible span of text in the paragraph** that answers the question!

- **Ημερομηνία:** 19 Οκτωβρίου 1512
- **Αριθμός:** 12
- **Άτομο:** Μανόλης Κουμπαράκης
- **Τοποθεσία:** Αθήνα
- **Οντότητα:** Ευρωπαϊκή Ένωση
- **Φράση ουσιαστικού:** καταστροφή ιδιοκτησίας
- **Φράση επιθέτου:** δεύτερο μεγαλύτερο
- **Φράση ρήματος:** επέστρεψε στην Γη
- **Πρόταση/Υποπρόταση:** για να αποφευχθεί ο πόλεμος
- **Άλλα:** προσεκτικά

Ερώτηση και Απάντηση, Παράδειγμα 1

Ερώτηση: Πως καλείται μερικές φορές ο κύκλος Rankine;

Τμήμα παραγράφου (**Απάντηση**):

> ...
> Ο κύκλος Rankine μερικές φορές αναφέρεται ως ένας **πρακτικός κύκλος Carnot**.
> ...

Ερώτηση και Απάντηση, Παράδειγμα 2

Ερώτηση: Ποια κυβερνητικά όργανα εχουν δικαίωμα βέτο;

Τμήμα παραγράφου (**Απάντηση**):

> ...
> **Το Ευρωπαϊκό Κοινοβούλιο και το Συμβούλιο της Ευρωπαϊκής Ένωσης** έχουν δικαίωμα τροποποίησης και βέτο κατά τη νομοθετική διαδικασία.
> ...

Ερώτηση και Απάντηση, Παράδειγμα 3

Ερώτηση: Ποιος ερευνητής του Shakespeare είναι επί του παρόντος στο Διδακτικό Ερευνητικό Προσωπικό;

Τμήμα παραγράφου (**Απάντηση**):

> …
> Το υπάρχον Διδακτικό Ερευνητικό Προσωπικό περιλαμβάνει τον ανθρωπολόγο Marshall Sahlins, … , τον μελετητή του Shakespeare **David Bevington**.
> …

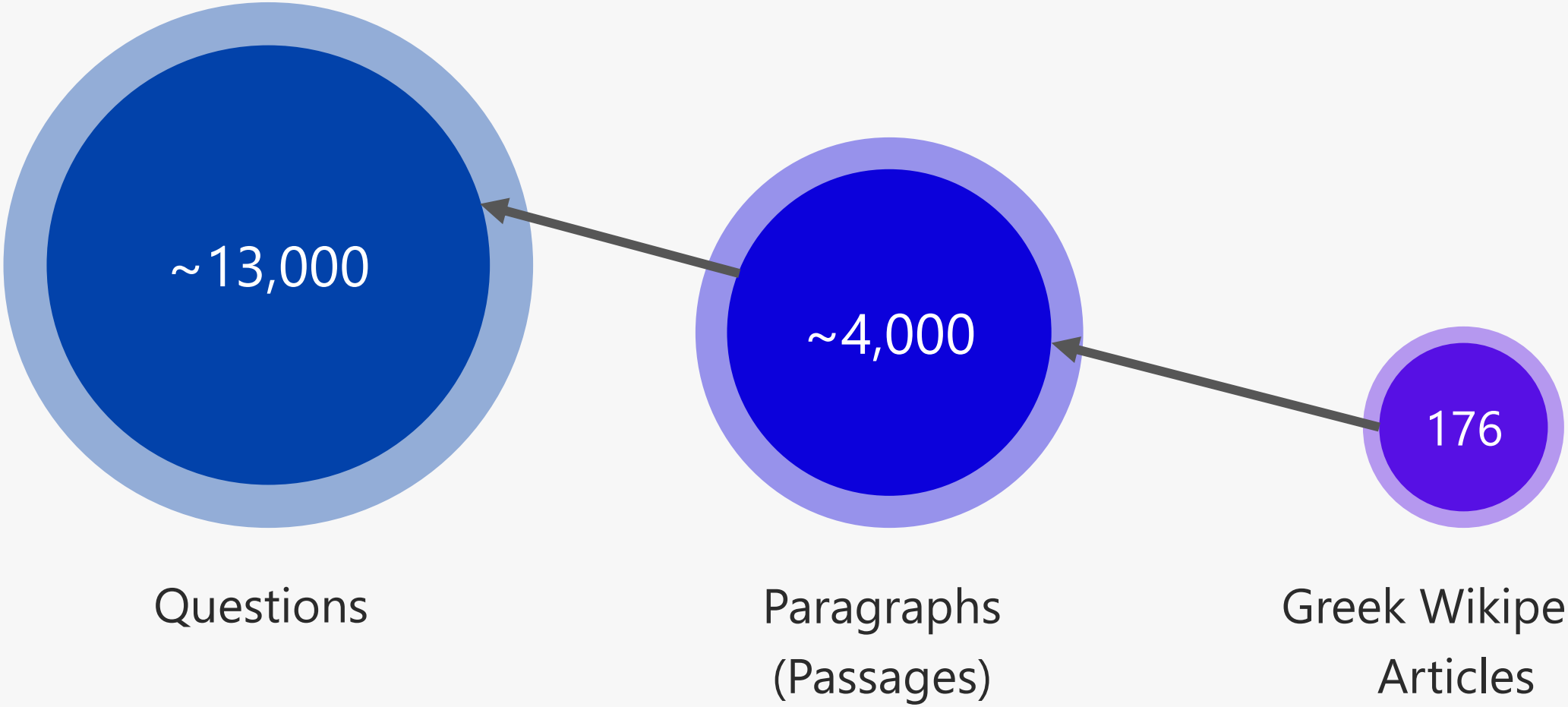
Ερώτηση και Απάντηση, Παράδειγμα 4

Ερώτηση: Τι συλλογή κατέχει η γκαλερί V&A Theatre & Performance;

Τμήμα παραγράφου (**Απάντηση**):

> ...
> Η γκαλερί V&A Theatre & Performance άνοιξε τον Μάρτιο του 2009.
> ...
> Αυτή η γκαλερί διατηρεί τη μεγαλύτερη εθνική συλλογή του Ηνωμένου Βασιλείου από **εκθέματα σχετικά με ζωντανές εμφανίσεις**.
> ...

Estimated Size of GreekQA Dataset



Questions?

Thank you for your contribution!

Stanford CS224N lecture slides
http://web.stanford.edu/class/cs224n/slides/cs224n-2022-lecture05-rnnlm.pdf
http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture10-QA.pdf

SQuAD1.0 (Rajpurkar '16)
https://arxiv.org/pdf/1606.05250.pdf