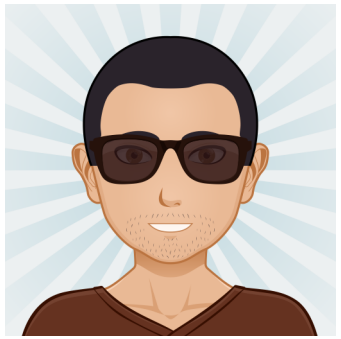
Global Maksimum AI Team

SadedeGel

sadedeGel Team

Dorukhan Afacan

- ML Development Coordination
- Literature survey
- BERT embedding models, aka Cluster Summarizers
- SadedeGel Dataset

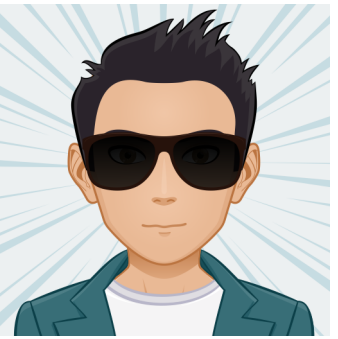


Murat Çakır

- sadedeGel Annotator
- sadedeGel Scraper
- sadedeGel Extension
- sadedeGel server

Askar Bozcan

- Literature survey
- Rouge1 Model
- Building Blocks (**bblock** module) development
- Summarizer OO design
- SadedeGel Dataset



Hüsnü Şensoy

- Commit Master
- Devops Engineering
- Baseline models
- SBD Implementation

sadedeGel Project Flow

Why don't we develop an **unsupervised extraction based Turkish news summarizer** based on BERT sentence embeddings ?
There are a few recent papers on the idea of using clustering those embeddings and …

Sure but before jumping into algorithmic details
1. What is our evaluation metric ?
2. What is a few baseline model scores based on this metric ?

☹ The problem is that we haven't seen a common turkish news corpus for summarization

I can develop a scraper collect necessary corpora
- Small corpus to prottype our idea
- Large corpus to continue development and share with other researchers.

SADEDEGEL Scraper

But we still have a problem because those will not be annotated for evaluation and how to define an **important sentence** ?

- Here is my definition:
  - A human annotator reads the news document and throws away some less important sentences in Pass 1
  - Then rereads what is left and throws away some more in Pass 2
  - This goes on until node more sentence left.
  - Latter a sentence is dropped, more its relative importance should be.
- If you are fine with the definition
  - Our evaluation metric is **Normalized Discounted Cumulative Gain**

Moreover we can devalop an annotator mimicing the behaviour in my definition.

SADEDEGEL Annotator

I don't want to be too negative but we are in trouble with the existing SBDs on collected datasets.

ML based SBD is not a rocket science. If you can annodate 100 sentences we can build one based on token features.

SADEDEGEL Library

Yeah… Let's do it !!!

Eeeh. I don't have much to do ?!?!

SADEDEGEL Chrome Extension

SadedeGel Scraper

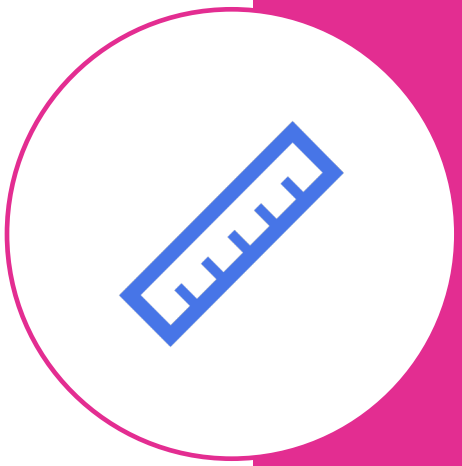
- A web scraper developed with **Scala** using open source **jsoup** library to meet the data requirements of SadedeGel library.
- It scrapes data from news websites and stores them as **.txt** files.
- We focused on the author pages of news websites rather than short frontpage news.
- Also picked certain authors from each news website that well fit with article format.
- sadegegel-scraper allows user defined scraper implementations by simply extending **NewsWebsite** trait.
    - sadedege-scraper Github README for more

Demo !!!

Normalized Discounted Cumulative Gain (N-DCG$_k$)

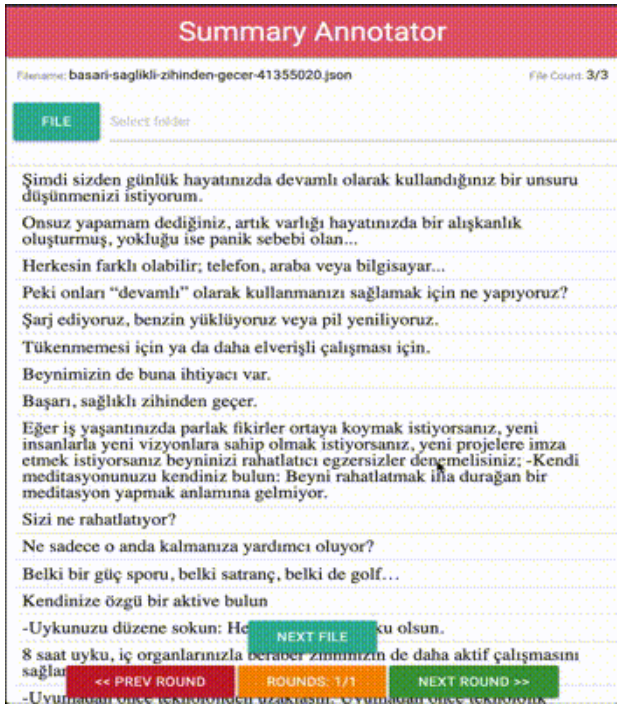
Sum sorted **k** selections normalized with sum sorted k most relevant

$$ndcg_k = \frac{\sum_i^k PickedSentence_i}{\sum_i^k RelevantSentence_i}$$



SadedeGel Annotator

Sadedegel Annotator is a cross platform annotation tool for extraction-based summarization implemented on electronJs.



Demo !!!

SadedeGel Sentence Boundary Detector (SBD)

- A few issues with online Turkish news corpus sentences
  - … bu da Demirel'in en büyük hatasıydı Tabi ki bu onun değerli bir devlet adamı…
  - Böyle futbol olmaz..! Gerçekten olmaz..

- Using rules defined in Speech and Language Processing, Second Edition we have created a ML based SBD
  - This SBD is shipped with sadedegel
  - User may choose to modify it

SBD Model	IoU Score (micro)	IoU Score (macro)
NLTKPunctTokenizer	0.7071	0.7343
RegexpSentenceTokenizer	0.6812	0.7224
MLBasedTokenizer	0.8881	0.8946

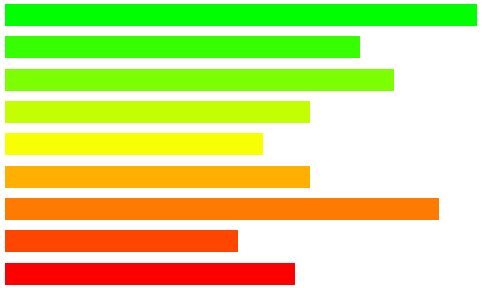
```
python3 -m sadedegel.tokenize build --help
Usage: __main__.py build [OPTIONS]

  Build a ML based SBD

Options:
  --help  Show this message and exit.
```

reproducable

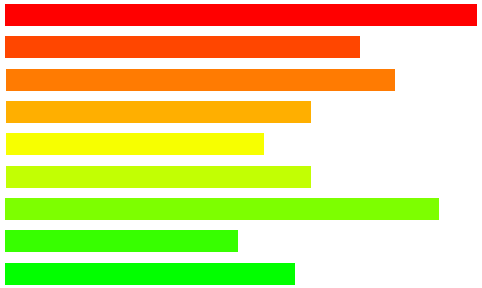
```
python3 -m sadedegel.tokenize evaluate --help
Usage: __main__.py evaluate [OPTIONS]

  Evaluate IoU metric for different SBD algorithms over our stock dataset.

Options:
  -v      verbosity
  --help  Show this message and exit.
```

SadedeGel Baseline Summarizer

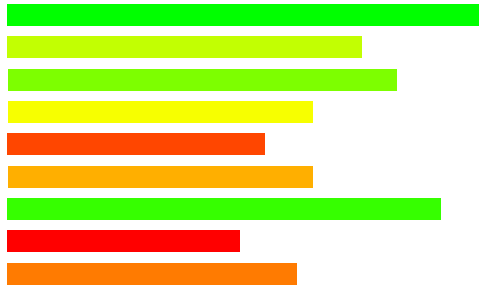
PositionSummarizer – First



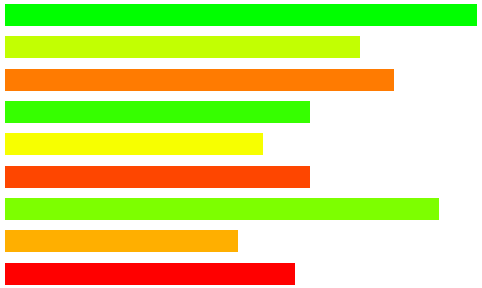
PositionSummarizer – Last



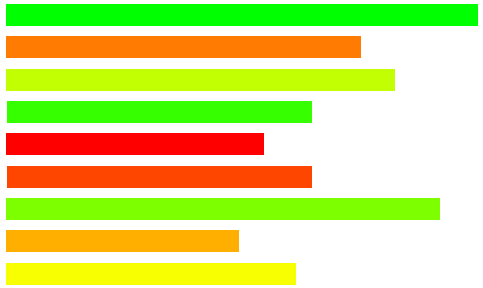
Length Summarizer (Token/Char)



Band Summarizer (k=3, forward)*



Random Summarize



* Model is not a part of sadedegel by the time this presentation is held.

Rouge1 Score By Example

SadedeGel Rouge1 Summarizer

1. Assume we have the following 3-sentences toy document

   Ali bakkaldan top aldı. Ali top oynadı. Ali eve gitti.

2. unigrams for each sentences

   [‘ali’, ‘bakkal’, ’##dan’, ‘top’, aldı’, ‘.’]
   [‘ali’, ‘top’, ‘oynadı’, ’.’]
   [‘ali’, ‘eve’, ‘gitti’,’.’]

3. Number of shared unigrams per sentences with all other sentences but itself

   2 (ali, top)
   2 (ali, top)
   1 (ali)

4. Rouge1

recall $\frac{num\_overlapping\_words}{total\_words\_in\_document}$

precision $\frac{num\_overlapping\_words}{total\_words\_in\_sentence}$

F1

$rouge1_{recall} = {}^{3}/_{8} \quad {}^{3}/_{10} \quad {}^{2}/_{10}$

$rouge1_{precision} = {}^{3}/_{6} \quad {}^{3}/_{4} \quad {}^{2}/_{4}$

Harmonic mean of Recall & Precision

SadedeGel Cluster Based Models

KMeansSummarizer (n_clusters = 2)

1. Get the BERT embedding for each sentences in document (768)
2. Cluster embeddings into **n_clusters**
3. Score each sentences based on its this to cluster center

AutoKMeansSummarizer (n_clustet_to_length = 0.05)

- Similar to **KMeansSummarizer**
- Number of clusters is defined based on document length. Longer documents have more clusters

DecomposedKMeansSummarizer (n_clusters = 2 , n_components = 48)

- Similar to **KMeansSummarizer**
- Before clustering, PCA decomposes embeddings into lower dimensions to obtain a denser vector representation.

How do we get those “BERT Embeddings” ?

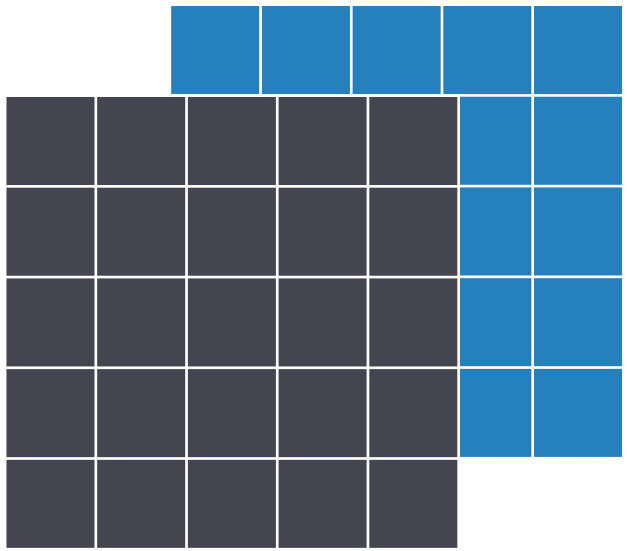
Sentence$_1$

[CLS]	Token$_1$	…	Token$_n$	[SEP]	[PAD]
1	1	…	1	1	0

Sentence$_2$

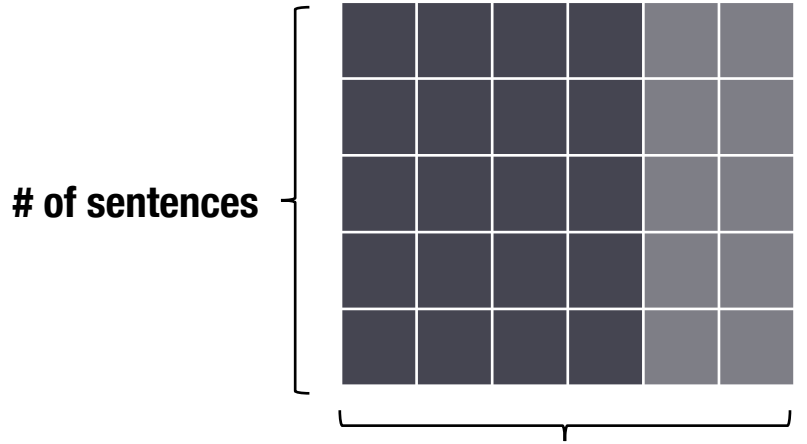
[CLS]	Token$_1$	…	[SEP]	[PAD]	[PAD]
1	1	…	1	0	0

Input Tensor for Document



Token IDs

Token Type IDs



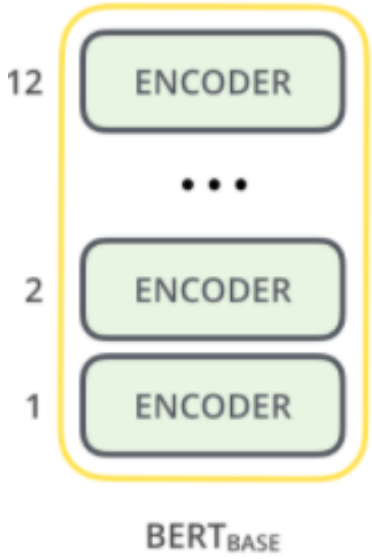
# of sentences

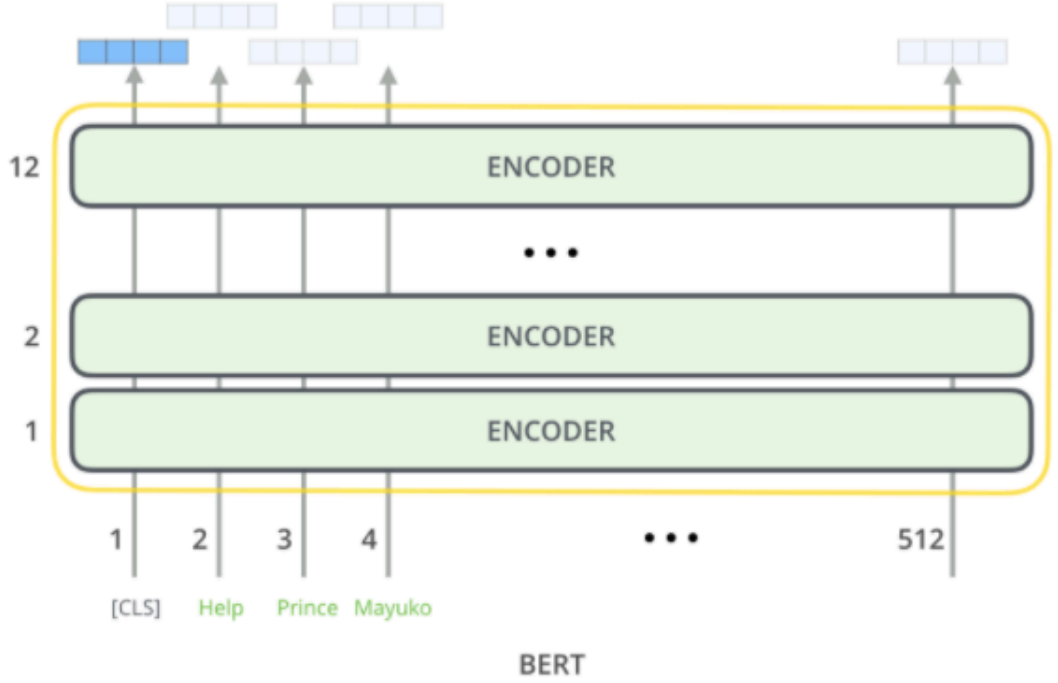
Max Tokens

Shorter sentences are **[PAD]**ed to match **Max Token** length

Just Enough BERT

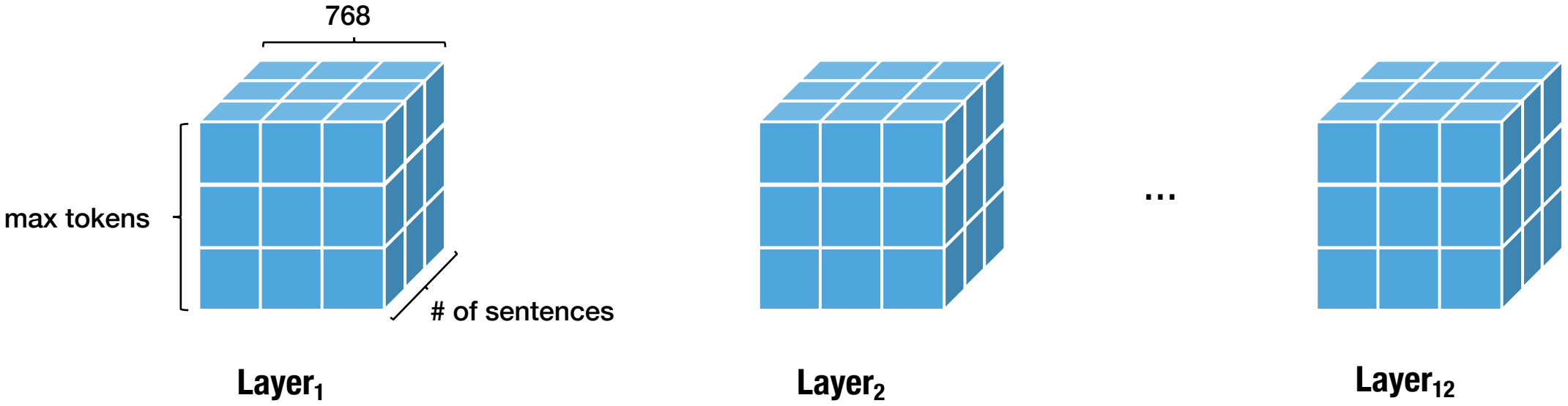
How do we get those “BERT Embeddings” ?





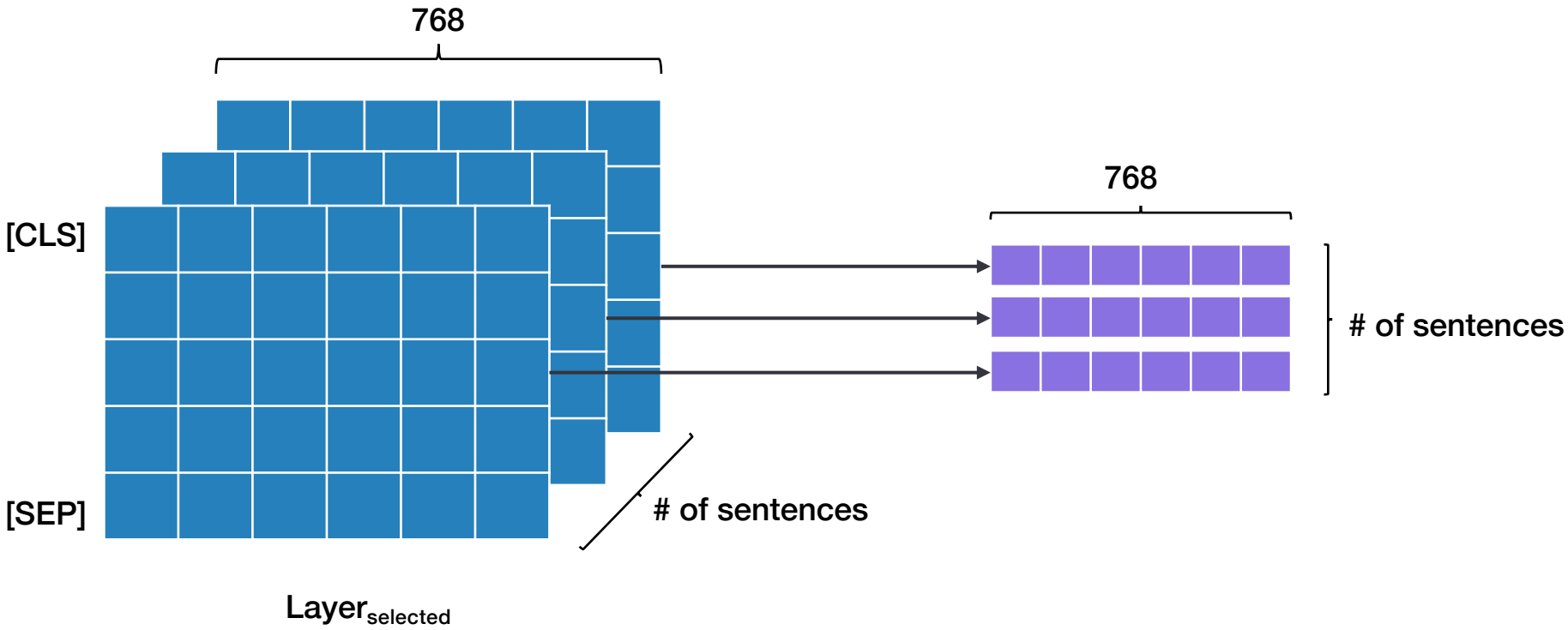
BERT Output per Document

How do we get those “BERT Embeddings” ?



Sentence Embedding with Average Pooling

How do we get those “BERT Embeddings” ?



Unsupervised Model on Self-Supervised Data*

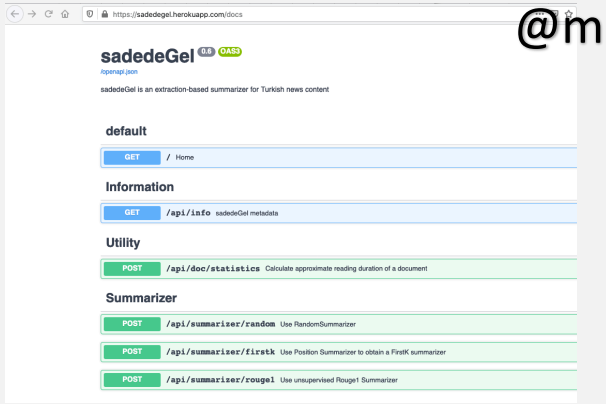
- Idea is to create a supervised model on document sentences
  - **X**: BERT embedding
  - **y**: Rouge1 score
- We believe that BERT embedding of a sentences is better than words themselves in obtaining a relevance score.

* Model is not a part of sadedegel by the time this presentation is held.

SadedeGel Summarizer Evaluation

Method	ndcg(k=0.1)	ndcg(k=0.5)	ndcg(k=0.8)
Random	0.5513	0.6502	0.7679
FirstK	0.5033	0.6154	0.7411
LastK	0.6048	0.6973	0.8013
Rouge1 (f1)	0.6727	0.7530	0.8447
Rouge1 (precision)	0.5293	0.6504	0.7745
Rouge1 (recall)	**0.6753**	0.7546	0.8452
Length (char)	0.6751	**0.7555**	0.8458
Length (token)	0.6753	0.7554	**0.8492**
KMeans	0.6569	0.7432	0.8336
AutoKMeansSummarizer	0.6576	0.7417	0.8324
DecomposedKMeansSummarizer	0.6550	0.7436	0.8331

SadedeGel Server



- HTTP based APIs with JSON payloads
- Backend of SadedeGel Chrome Extension deployed on Heroku free tier

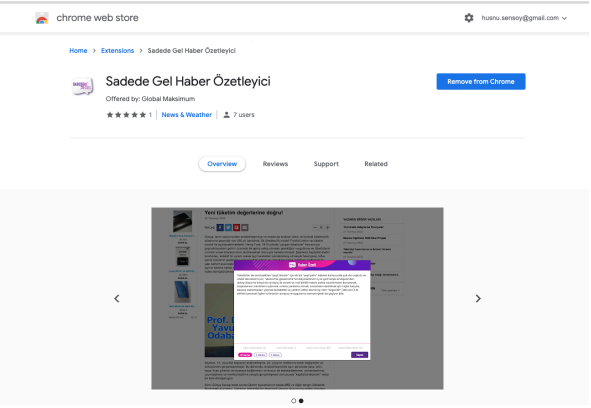
```
~/cod/sadedegel on develop !1 ?10 python3 -m sadedegel.server --help
Usage: __main__.py [OPTIONS]

  Span a sadedeGel http server instance.

Options:
  -h, --host TEXT           Hostname
  --log-level [debug|info]  Logging Level
  --reload                  enable/disable auto reload for development.
  --port INTEGER            Port
  --help                    Show this message and exit.
```

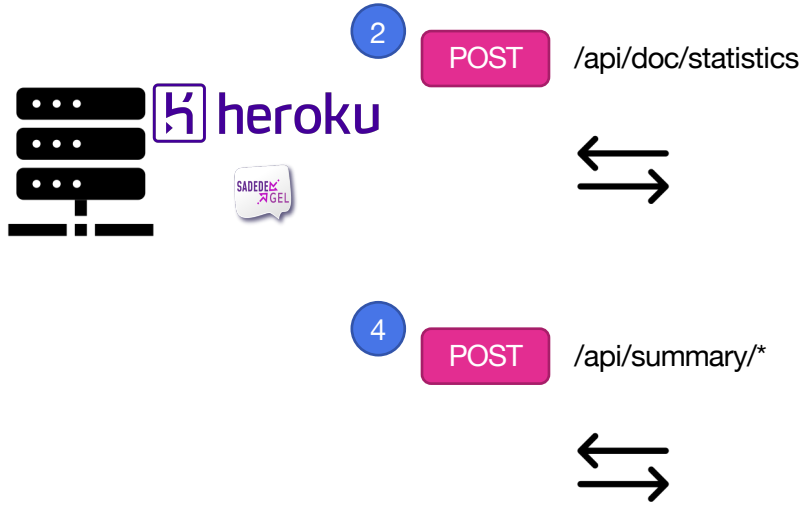
```
~/cod/sadedegel on develop !1 ?10 python3 -m sadedegel.server
INFO:     Started server process [87987]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
```

Demo !!!

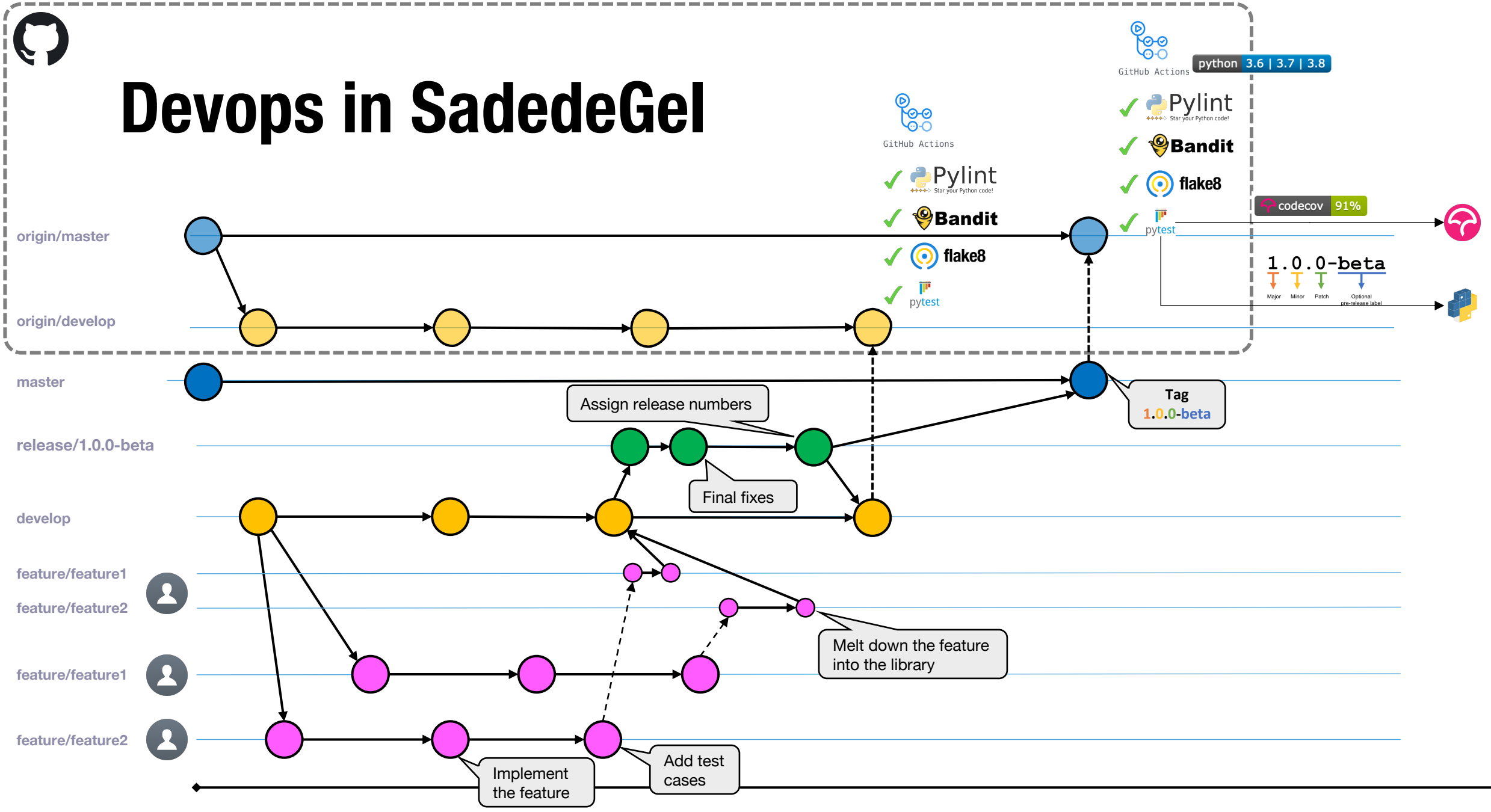
SadedeGel Chrome Extension

- sadedeGel Chrome Extension is developed to make sadedeGel accessible by the end users
- It allows you the summarize articles on authors page of supported news websites.
  - We have developed necessary customizations to auto detect article text within well known news websites
    - hurriyet.com.tr
    - milliyet.com.tr
    - sozcu.com.tr
    - haberturk.com
    - sabah.com.tr
- Also you can install it in Chrome Developer Mode.
  - Details on [sadedegel-chrome-extension Github]





Devops in SadedeGel

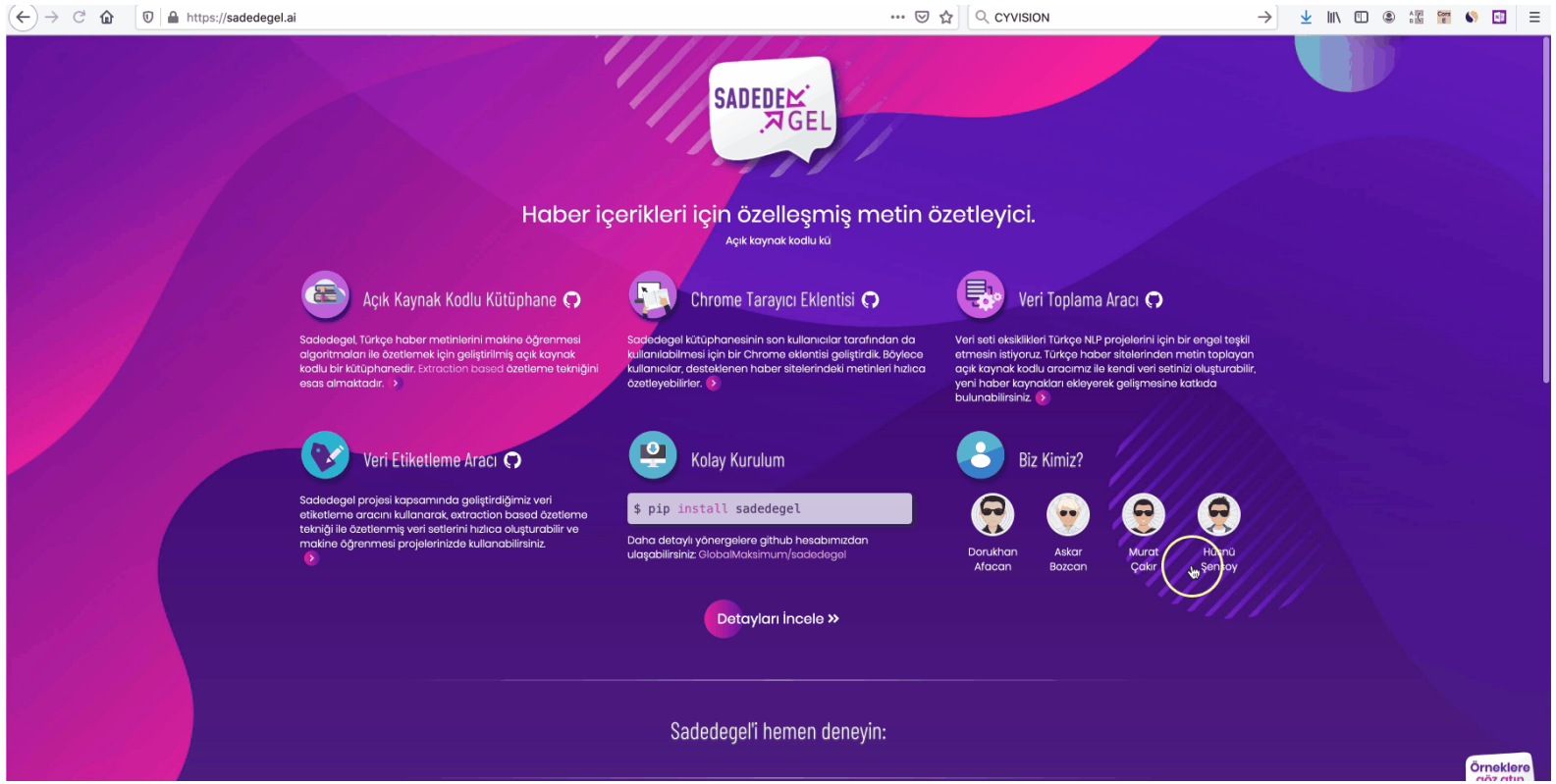


Q & A

SadedeGel

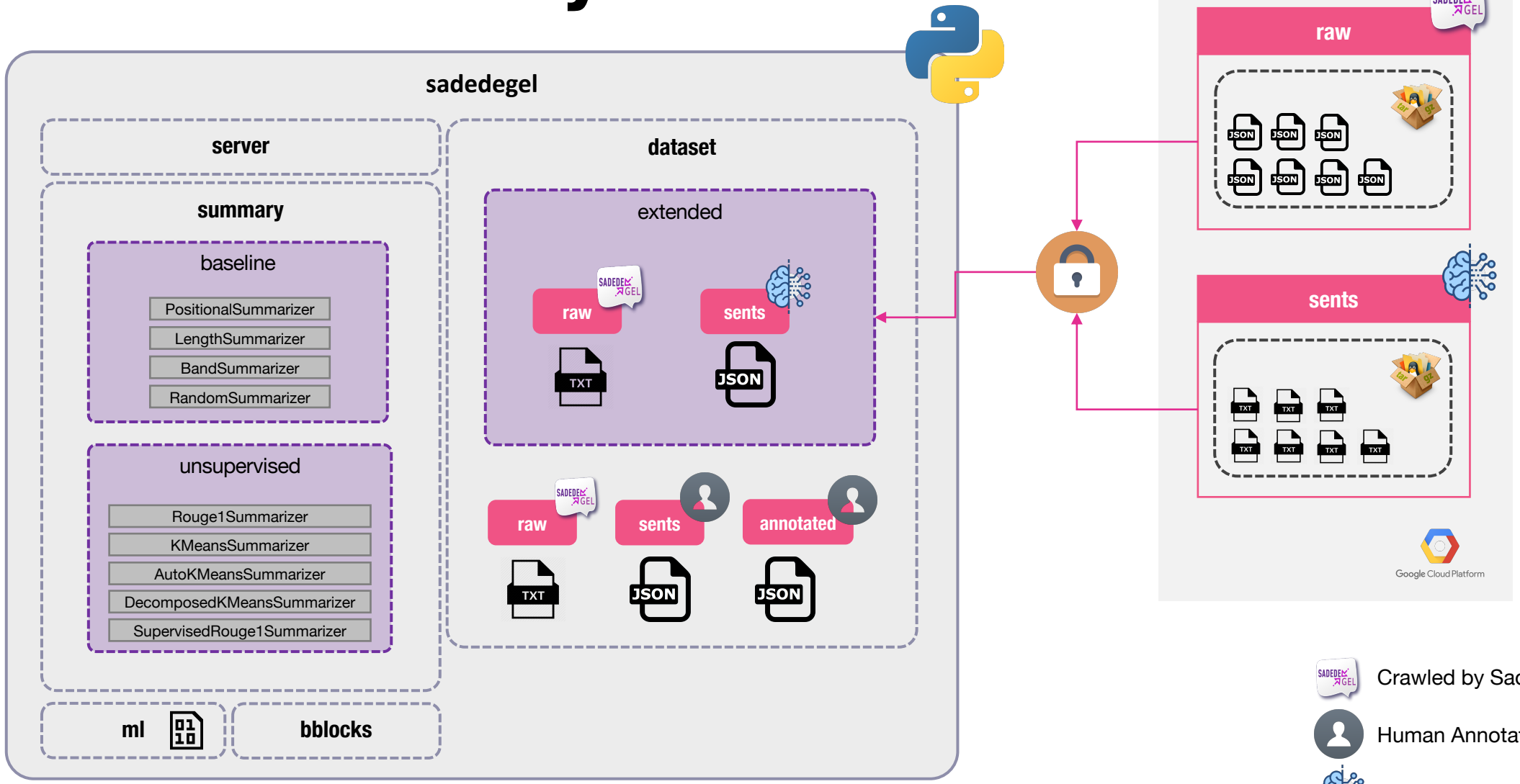
Backup Slides

Binder Integration

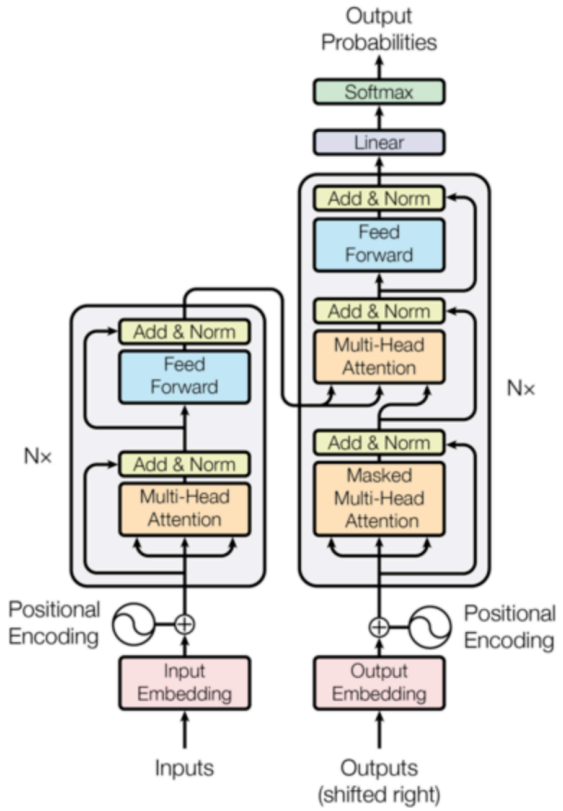


- SadedeGel also has binder integration for
  - Starting a **Jupyter** notebook without any installation on your client machine
  - Embedded examples on sadedegel.ai

SadedeGel Libray



Bert Specs

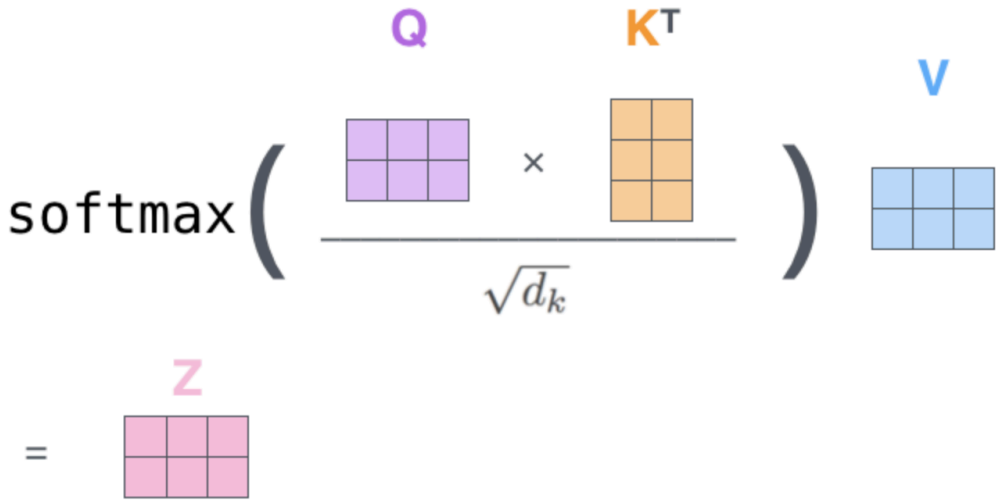


Modular Structure for Transfer Learning Tasks

$$PE(pos = 2) = (sin(2), cos(2), sin(1.88), cos(1.88), ..., sin(0.62), cos(0.62), ...)$$

First 40 values of the Positional Encoding vector of "bird", at position 2.

Positional Encoding



The self-attention calculation in matrix form

Attention Mechanism

PEGASUS

- Inspired Sadedegel's Rouge1 summarizer
- Abstractive summarizer
- Does not need large amount of labelled data