assignment_1_A

April 6, 2025

1 ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ - Ε 1

1.1 A. Tokens, Types, Zipf's Law

Μ : Ε Φ Γ

Σ : Ι Κ

Ε : 2025-03-19 | v.0.0.1

1.2 Β 1: Ε Β /Ε

[59]:

```
%pip install nltk spacy transformers matplotlib pandas plotly
!python -m spacy download en_core_web_sm
%pip install --upgrade jupyter ipywidgets
```

```
Requirement already satisfied: nltk in ./.conda/lib/python3.11/site-packages
(3.9.1)
Requirement already satisfied: spacy in ./.conda/lib/python3.11/site-packages
(3.8.4)
Requirement already satisfied: transformers in ./.conda/lib/python3.11/site-
packages (4.49.0)
Requirement already satisfied: matplotlib in ./.conda/lib/python3.11/site-
packages (3.10.1)
Requirement already satisfied: pandas in ./.conda/lib/python3.11/site-packages
(2.2.3)
Requirement already satisfied: plotly in ./.conda/lib/python3.11/site-packages
(6.0.1)
Requirement already satisfied: click in ./.conda/lib/python3.11/site-packages
(from nltk) (8.1.8)
Requirement already satisfied: joblib in ./.conda/lib/python3.11/site-packages
(from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in ./.conda/lib/python3.11/site-
packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in ./.conda/lib/python3.11/site-packages
(from nltk) (4.67.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in
./.conda/lib/python3.11/site-packages (from spacy) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in
./.conda/lib/python3.11/site-packages (from spacy) (1.0.5)
```

```
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
./.conda/lib/python3.11/site-packages (from spacy) (1.0.12)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
./.conda/lib/python3.11/site-packages (from spacy) (2.0.11)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
./.conda/lib/python3.11/site-packages (from spacy) (3.0.9)
Requirement already satisfied: thinc<8.4.0,>=8.3.4 in
./.conda/lib/python3.11/site-packages (from spacy) (8.3.4)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in
./.conda/lib/python3.11/site-packages (from spacy) (1.1.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in
./.conda/lib/python3.11/site-packages (from spacy) (2.5.1)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in
./.conda/lib/python3.11/site-packages (from spacy) (2.0.10)
Requirement already satisfied: weasel<0.5.0,>=0.1.0 in
./.conda/lib/python3.11/site-packages (from spacy) (0.4.1)
Requirement already satisfied: typer<1.0.0,>=0.3.0 in
./.conda/lib/python3.11/site-packages (from spacy) (0.15.2)
Requirement already satisfied: numpy>=1.19.0 in ./.conda/lib/python3.11/site-
packages (from spacy) (2.2.4)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in
./.conda/lib/python3.11/site-packages (from spacy) (2.32.3)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in
./.conda/lib/python3.11/site-packages (from spacy) (2.10.6)
Requirement already satisfied: jinja2 in ./.conda/lib/python3.11/site-packages
(from spacy) (3.1.6)
Requirement already satisfied: setuptools in ./.conda/lib/python3.11/site-
packages (from spacy) (75.8.0)
Requirement already satisfied: packaging>=20.0 in ./.conda/lib/python3.11/site-
packages (from spacy) (24.2)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in
./.conda/lib/python3.11/site-packages (from spacy) (3.5.0)
Requirement already satisfied: filelock in ./.conda/lib/python3.11/site-packages
(from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.26.0 in
./.conda/lib/python3.11/site-packages (from transformers) (0.29.3)
Requirement already satisfied: pyyaml>=5.1 in ./.conda/lib/python3.11/site-
packages (from transformers) (6.0.2)
Requirement already satisfied: tokenizers<0.22,>=0.21 in
./.conda/lib/python3.11/site-packages (from transformers) (0.21.1)
Requirement already satisfied: safetensors>=0.4.1 in
./.conda/lib/python3.11/site-packages (from transformers) (0.5.3)
Requirement already satisfied: contourpy>=1.0.1 in ./.conda/lib/python3.11/site-
packages (from matplotlib) (1.3.1)
Requirement already satisfied: cycler>=0.10 in ./.conda/lib/python3.11/site-
packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
./.conda/lib/python3.11/site-packages (from matplotlib) (4.56.0)
```

```
Requirement already satisfied: kiwisolver>=1.3.1 in
./.conda/lib/python3.11/site-packages (from matplotlib) (1.4.8)
Requirement already satisfied: pillow>=8 in ./.conda/lib/python3.11/site-
packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in ./.conda/lib/python3.11/site-
packages (from matplotlib) (3.2.1)
Requirement already satisfied: python-dateutil>=2.7 in
./.conda/lib/python3.11/site-packages (from matplotlib) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in ./.conda/lib/python3.11/site-
packages (from pandas) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in ./.conda/lib/python3.11/site-
packages (from pandas) (2025.1)
Requirement already satisfied: narwhals>=1.15.1 in ./.conda/lib/python3.11/site-
packages (from plotly) (1.31.0)
Requirement already satisfied: fsspec>=2023.5.0 in ./.conda/lib/python3.11/site-
packages (from huggingface-hub<1.0,>=0.26.0->transformers) (2025.3.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
./.conda/lib/python3.11/site-packages (from huggingface-
hub<1.0,>=0.26.0->transformers) (4.12.2)
Requirement already satisfied: language-data>=1.2 in
./.conda/lib/python3.11/site-packages (from langcodes<4.0.0,>=3.2.0->spacy)
(1.3.0)
Requirement already satisfied: annotated-types>=0.6.0 in
./.conda/lib/python3.11/site-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (0.7.0)
Requirement already satisfied: pydantic-core==2.27.2 in
./.conda/lib/python3.11/site-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (2.27.2)
Requirement already satisfied: six>=1.5 in ./.conda/lib/python3.11/site-packages
(from python-dateutil>=2.7->matplotlib) (1.17.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
./.conda/lib/python3.11/site-packages (from requests<3.0.0,>=2.13.0->spacy)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in ./.conda/lib/python3.11/site-
packages (from requests<3.0.0,>=2.13.0->spacy) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
./.conda/lib/python3.11/site-packages (from requests<3.0.0,>=2.13.0->spacy)
(2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
./.conda/lib/python3.11/site-packages (from requests<3.0.0,>=2.13.0->spacy)
(2025.1.31)
Requirement already satisfied: blis<1.3.0,>=1.2.0 in
./.conda/lib/python3.11/site-packages (from thinc<8.4.0,>=8.3.4->spacy) (1.2.0)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in
./.conda/lib/python3.11/site-packages (from thinc<8.4.0,>=8.3.4->spacy) (0.1.5)
Requirement already satisfied: shellingham>=1.3.0 in
./.conda/lib/python3.11/site-packages (from typer<1.0.0,>=0.3.0->spacy) (1.5.4)
Requirement already satisfied: rich>=10.11.0 in ./.conda/lib/python3.11/site-
```

```
packages (from typer<1.0.0,>=0.3.0->spacy) (13.9.4)
Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in
./.conda/lib/python3.11/site-packages (from weasel<0.5.0,>=0.1.0->spacy)
(0.21.0)
Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in
./.conda/lib/python3.11/site-packages (from weasel<0.5.0,>=0.1.0->spacy) (7.1.0)
Requirement already satisfied: MarkupSafe>=2.0 in ./.conda/lib/python3.11/site-
packages (from jinja2->spacy) (3.0.2)
Requirement already satisfied: marisa-trie>=1.1.0 in
./.conda/lib/python3.11/site-packages (from language-
data>=1.2->langcodes<4.0.0,>=3.2.0->spacy) (1.2.1)
Requirement already satisfied: markdown-it-py>=2.2.0 in
./.conda/lib/python3.11/site-packages (from
rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
./.conda/lib/python3.11/site-packages (from
rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy) (2.19.1)
Requirement already satisfied: wrapt in ./.conda/lib/python3.11/site-packages
(from smart-open<8.0.0,>=5.2.1->weasel<0.5.0,>=0.1.0->spacy) (1.17.2)
Requirement already satisfied: mdurl~=0.1 in ./.conda/lib/python3.11/site-
packages (from markdown-it-py>=2.2.0->rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy)
(0.1.2)
Note: you may need to restart the kernel to use updated packages.
Collecting en-core-web-sm==3.8.0
  Downloading https://github.com/explosion/spacy-
models/releases/download/en_core_web_sm-3.8.0/en_core_web_sm-3.8.0-py3-none-
any.whl (12.8 MB)
                           12.8/12.8 MB
4.1 MB/s eta 0:00:00a 0:00:01
 Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
Requirement already satisfied: jupyter in ./.conda/lib/python3.11/site-packages
(1.1.1)
Requirement already satisfied: ipywidgets in ./.conda/lib/python3.11/site-
packages (8.1.5)
Requirement already satisfied: notebook in ./.conda/lib/python3.11/site-packages
(from jupyter) (7.3.3)
Requirement already satisfied: jupyter-console in ./.conda/lib/python3.11/site-
packages (from jupyter) (6.6.3)
Requirement already satisfied: nbconvert in ./.conda/lib/python3.11/site-
packages (from jupyter) (7.16.6)
Requirement already satisfied: ipykernel in ./.conda/lib/python3.11/site-
packages (from jupyter) (6.29.5)
Requirement already satisfied: jupyterlab in ./.conda/lib/python3.11/site-
packages (from jupyter) (4.3.6)
Requirement already satisfied: comm>=0.1.3 in ./.conda/lib/python3.11/site-
packages (from ipywidgets) (0.2.2)
Requirement already satisfied: ipython>=6.1.0 in ./.conda/lib/python3.11/site-
```

```
packages (from ipywidgets) (9.0.2)
Requirement already satisfied: traitlets>=4.3.1 in ./.conda/lib/python3.11/site-
packages (from ipywidgets) (5.14.3)
Requirement already satisfied: widgetsnbextension~=4.0.12 in
./.conda/lib/python3.11/site-packages (from ipywidgets) (4.0.13)
Requirement already satisfied: jupyterlab-widgets~=3.0.12 in
./.conda/lib/python3.11/site-packages (from ipywidgets) (3.0.13)
Requirement already satisfied: decorator in ./.conda/lib/python3.11/site-
packages (from ipython>=6.1.0->ipywidgets) (5.2.1)
Requirement already satisfied: ipython-pygments-lexers in
./.conda/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (1.1.1)
Requirement already satisfied: jedi>=0.16 in ./.conda/lib/python3.11/site-
packages (from ipython>=6.1.0->ipywidgets) (0.19.2)
Requirement already satisfied: matplotlib-inline in
./.conda/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (0.1.7)
Requirement already satisfied: pexpect>4.3 in ./.conda/lib/python3.11/site-
packages (from ipython>=6.1.0->ipywidgets) (4.9.0)
Requirement already satisfied: prompt_toolkit<3.1.0,>=3.0.41 in
./.conda/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (3.0.50)
Requirement already satisfied: pygments>=2.4.0 in ./.conda/lib/python3.11/site-
packages (from ipython>=6.1.0->ipywidgets) (2.19.1)
Requirement already satisfied: stack_data in ./.conda/lib/python3.11/site-
packages (from ipython>=6.1.0->ipywidgets) (0.6.3)
Requirement already satisfied: typing_extensions>=4.6 in
./.conda/lib/python3.11/site-packages (from ipython>=6.1.0->ipywidgets) (4.12.2)
Requirement already satisfied: appnope in ./.conda/lib/python3.11/site-packages
(from ipykernel->jupyter) (0.1.4)
Requirement already satisfied: debugpy>=1.6.5 in ./.conda/lib/python3.11/site-
packages (from ipykernel->jupyter) (1.8.13)
Requirement already satisfied: jupyter-client>=6.1.12 in
./.conda/lib/python3.11/site-packages (from ipykernel->jupyter) (8.6.3)
Requirement already satisfied: jupyter-core!=5.0.*,>=4.12 in
./.conda/lib/python3.11/site-packages (from ipykernel->jupyter) (5.7.2)
Requirement already satisfied: nest-asyncio in ./.conda/lib/python3.11/site-
packages (from ipykernel->jupyter) (1.6.0)
Requirement already satisfied: packaging in ./.conda/lib/python3.11/site-
packages (from ipykernel->jupyter) (24.2)
Requirement already satisfied: psutil in ./.conda/lib/python3.11/site-packages
(from ipykernel->jupyter) (7.0.0)
Requirement already satisfied: pyzmq>=24 in ./.conda/lib/python3.11/site-
packages (from ipykernel->jupyter) (26.3.0)
Requirement already satisfied: tornado>=6.1 in ./.conda/lib/python3.11/site-
packages (from ipykernel->jupyter) (6.4.2)
Requirement already satisfied: async-lru>=1.0.0 in ./.conda/lib/python3.11/site-
packages (from jupyterlab->jupyter) (2.0.5)
Requirement already satisfied: httpx>=0.25.0 in ./.conda/lib/python3.11/site-
packages (from jupyterlab->jupyter) (0.28.1)
Requirement already satisfied: jinja2>=3.0.3 in ./.conda/lib/python3.11/site-
```

```
packages (from jupyterlab->jupyter) (3.1.6)
Requirement already satisfied: jupyter-lsp>=2.0.0 in
./.conda/lib/python3.11/site-packages (from jupyterlab->jupyter) (2.2.5)
Requirement already satisfied: jupyter-server<3,>=2.4.0 in
./.conda/lib/python3.11/site-packages (from jupyterlab->jupyter) (2.15.0)
Requirement already satisfied: jupyterlab-server<3,>=2.27.1 in
./.conda/lib/python3.11/site-packages (from jupyterlab->jupyter) (2.27.3)
Requirement already satisfied: notebook-shim>=0.2 in
./.conda/lib/python3.11/site-packages (from jupyterlab->jupyter) (0.2.4)
Requirement already satisfied: setuptools>=40.8.0 in
./.conda/lib/python3.11/site-packages (from jupyterlab->jupyter) (75.8.0)
Requirement already satisfied: beautifulsoup4 in ./.conda/lib/python3.11/site-
packages (from nbconvert->jupyter) (4.13.3)
Requirement already satisfied: bleach!=5.0.0 in ./.conda/lib/python3.11/site-
packages (from bleach[css]!=5.0.0->nbconvert->jupyter) (6.2.0)
Requirement already satisfied: defusedxml in ./.conda/lib/python3.11/site-
packages (from nbconvert->jupyter) (0.7.1)
Requirement already satisfied: jupyterlab-pygments in
./.conda/lib/python3.11/site-packages (from nbconvert->jupyter) (0.3.0)
Requirement already satisfied: markupsafe>=2.0 in ./.conda/lib/python3.11/site-
packages (from nbconvert->jupyter) (3.0.2)
Requirement already satisfied: mistune<4,>=2.0.3 in
./.conda/lib/python3.11/site-packages (from nbconvert->jupyter) (3.1.3)
Requirement already satisfied: nbclient>=0.5.0 in ./.conda/lib/python3.11/site-
packages (from nbconvert->jupyter) (0.10.2)
Requirement already satisfied: nbformat>=5.7 in ./.conda/lib/python3.11/site-
packages (from nbconvert->jupyter) (5.10.4)
Requirement already satisfied: pandocfilters>=1.4.1 in
./.conda/lib/python3.11/site-packages (from nbconvert->jupyter) (1.5.1)
Requirement already satisfied: webencodings in ./.conda/lib/python3.11/site-
packages (from bleach!=5.0.0->bleach[css]!=5.0.0->nbconvert->jupyter) (0.5.1)
Requirement already satisfied: tinycss2<1.5,>=1.1.0 in
./.conda/lib/python3.11/site-packages (from
bleach[css]!=5.0.0->nbconvert->jupyter) (1.4.0)
Requirement already satisfied: anyio in ./.conda/lib/python3.11/site-packages
(from httpx>=0.25.0->jupyterlab->jupyter) (4.9.0)
Requirement already satisfied: certifi in ./.conda/lib/python3.11/site-packages
(from httpx>=0.25.0->jupyterlab->jupyter) (2025.1.31)
Requirement already satisfied: httpcore==1.* in ./.conda/lib/python3.11/site-
packages (from httpx>=0.25.0->jupyterlab->jupyter) (1.0.7)
Requirement already satisfied: idna in ./.conda/lib/python3.11/site-packages
(from httpx>=0.25.0->jupyterlab->jupyter) (3.10)
Requirement already satisfied: h11<0.15,>=0.13 in ./.conda/lib/python3.11/site-
packages (from httpcore==1.*->httpx>=0.25.0->jupyterlab->jupyter) (0.14.0)
Requirement already satisfied: parso<0.9.0,>=0.8.4 in
./.conda/lib/python3.11/site-packages (from
jedi>=0.16->ipython>=6.1.0->ipywidgets) (0.8.4)
Requirement already satisfied: python-dateutil>=2.8.2 in
```

```
./.conda/lib/python3.11/site-packages (from jupyter-
client>=6.1.12->ipykernel->jupyter) (2.9.0.post0)
Requirement already satisfied: platformdirs>=2.5 in
./.conda/lib/python3.11/site-packages (from jupyter-
core!=5.0.*,>=4.12->ipykernel->jupyter) (4.3.6)
Requirement already satisfied: argon2-cffi>=21.1 in
./.conda/lib/python3.11/site-packages (from jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (23.1.0)
Requirement already satisfied: jupyter-events>=0.11.0 in
./.conda/lib/python3.11/site-packages (from jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (0.12.0)
Requirement already satisfied: jupyter-server-terminals>=0.4.4 in
./.conda/lib/python3.11/site-packages (from jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (0.5.3)
Requirement already satisfied: overrides>=5.0 in ./.conda/lib/python3.11/site-
packages (from jupyter-server<3,>=2.4.0->jupyterlab->jupyter) (7.7.0)
Requirement already satisfied: prometheus-client>=0.9 in
./.conda/lib/python3.11/site-packages (from jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (0.21.1)
Requirement already satisfied: send2trash>=1.8.2 in
./.conda/lib/python3.11/site-packages (from jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (1.8.3)
Requirement already satisfied: terminado>=0.8.3 in ./.conda/lib/python3.11/site-
packages (from jupyter-server<3,>=2.4.0->jupyterlab->jupyter) (0.18.1)
Requirement already satisfied: websocket-client>=1.7 in
./.conda/lib/python3.11/site-packages (from jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (1.8.0)
Requirement already satisfied: babel>=2.10 in ./.conda/lib/python3.11/site-
packages (from jupyterlab-server<3,>=2.27.1->jupyterlab->jupyter) (2.17.0)
Requirement already satisfied: json5>=0.9.0 in ./.conda/lib/python3.11/site-
packages (from jupyterlab-server<3,>=2.27.1->jupyterlab->jupyter) (0.10.0)
Requirement already satisfied: jsonschema>=4.18.0 in
./.conda/lib/python3.11/site-packages (from jupyterlab-
server<3,>=2.27.1->jupyterlab->jupyter) (4.23.0)
Requirement already satisfied: requests>=2.31 in ./.conda/lib/python3.11/site-
packages (from jupyterlab-server<3,>=2.27.1->jupyterlab->jupyter) (2.32.3)
Requirement already satisfied: fastjsonschema>=2.15 in
./.conda/lib/python3.11/site-packages (from nbformat>=5.7->nbconvert->jupyter)
(2.21.1)
Requirement already satisfied: ptyprocess>=0.5 in ./.conda/lib/python3.11/site-
packages (from pexpect>4.3->ipython>=6.1.0->ipywidgets) (0.7.0)
Requirement already satisfied: wcwidth in ./.conda/lib/python3.11/site-packages
(from prompt_toolkit<3.1.0,>=3.0.41->ipython>=6.1.0->ipywidgets) (0.2.13)
Requirement already satisfied: soupsieve>1.2 in ./.conda/lib/python3.11/site-
packages (from beautifulsoup4->nbconvert->jupyter) (2.6)
Requirement already satisfied: executing>=1.2.0 in ./.conda/lib/python3.11/site-
packages (from stack_data->ipython>=6.1.0->ipywidgets) (2.2.0)
Requirement already satisfied: asttokens>=2.1.0 in ./.conda/lib/python3.11/site-
```

```
packages (from stack_data->ipython>=6.1.0->ipywidgets) (3.0.0)
Requirement already satisfied: pure-eval in ./.conda/lib/python3.11/site-
packages (from stack_data->ipython>=6.1.0->ipywidgets) (0.2.3)
Requirement already satisfied: sniffio>=1.1 in ./.conda/lib/python3.11/site-
packages (from anyio->httpx>=0.25.0->jupyterlab->jupyter) (1.3.1)
Requirement already satisfied: argon2-cffi-bindings in
./.conda/lib/python3.11/site-packages (from argon2-cffi>=21.1->jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (21.2.0)
Requirement already satisfied: attrs>=22.2.0 in ./.conda/lib/python3.11/site-
packages (from jsonschema>=4.18.0->jupyterlab-
server<3,>=2.27.1->jupyterlab->jupyter) (25.3.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
./.conda/lib/python3.11/site-packages (from jsonschema>=4.18.0->jupyterlab-
server<3,>=2.27.1->jupyterlab->jupyter) (2024.10.1)
Requirement already satisfied: referencing>=0.28.4 in
./.conda/lib/python3.11/site-packages (from jsonschema>=4.18.0->jupyterlab-
server<3,>=2.27.1->jupyterlab->jupyter) (0.36.2)
Requirement already satisfied: rpds-py>=0.7.1 in ./.conda/lib/python3.11/site-
packages (from jsonschema>=4.18.0->jupyterlab-
server<3,>=2.27.1->jupyterlab->jupyter) (0.23.1)
Requirement already satisfied: python-json-logger>=2.0.4 in
./.conda/lib/python3.11/site-packages (from jupyter-events>=0.11.0->jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (3.3.0)
Requirement already satisfied: pyyaml>=5.3 in ./.conda/lib/python3.11/site-
packages (from jupyter-events>=0.11.0->jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (6.0.2)
Requirement already satisfied: rfc3339-validator in
./.conda/lib/python3.11/site-packages (from jupyter-events>=0.11.0->jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (0.1.4)
Requirement already satisfied: rfc3986-validator>=0.1.1 in
./.conda/lib/python3.11/site-packages (from jupyter-events>=0.11.0->jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (0.1.1)
Requirement already satisfied: six>=1.5 in ./.conda/lib/python3.11/site-packages
(from python-dateutil>=2.8.2->jupyter-client>=6.1.12->ipykernel->jupyter)
(1.17.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
./.conda/lib/python3.11/site-packages (from requests>=2.31->jupyterlab-
server<3,>=2.27.1->jupyterlab->jupyter) (3.4.1)
Requirement already satisfied: urllib3<3,>=1.21.1 in
./.conda/lib/python3.11/site-packages (from requests>=2.31->jupyterlab-
server<3,>=2.27.1->jupyterlab->jupyter) (2.3.0)
Requirement already satisfied: fqdn in ./.conda/lib/python3.11/site-packages
(from jsonschema[format-nongpl]>=4.18.0->jupyter-events>=0.11.0->jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (1.5.1)
Requirement already satisfied: isoduration in ./.conda/lib/python3.11/site-
packages (from jsonschema[format-nongpl]>=4.18.0->jupyter-
events>=0.11.0->jupyter-server<3,>=2.4.0->jupyterlab->jupyter) (20.11.0)
Requirement already satisfied: jsonpointer>1.13 in ./.conda/lib/python3.11/site-
```

```
packages (from jsonschema[format-nongpl]>=4.18.0->jupyter-
events>=0.11.0->jupyter-server<3,>=2.4.0->jupyterlab->jupyter) (3.0.0)
Requirement already satisfied: uri-template in ./.conda/lib/python3.11/site-
packages (from jsonschema[format-nongpl]>=4.18.0->jupyter-
events>=0.11.0->jupyter-server<3,>=2.4.0->jupyterlab->jupyter) (1.3.0)
Requirement already satisfied: webcolors>=24.6.0 in
./.conda/lib/python3.11/site-packages (from jsonschema[format-
nongpl]>=4.18.0->jupyter-events>=0.11.0->jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (24.11.1)
Requirement already satisfied: cffi>=1.0.1 in ./.conda/lib/python3.11/site-
packages (from argon2-cffi-bindings->argon2-cffi>=21.1->jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (1.17.1)
Requirement already satisfied: pycparser in ./.conda/lib/python3.11/site-
packages (from cffi>=1.0.1->argon2-cffi-bindings->argon2-cffi>=21.1->jupyter-
server<3,>=2.4.0->jupyterlab->jupyter) (2.22)
Requirement already satisfied: arrow>=0.15.0 in ./.conda/lib/python3.11/site-
packages (from isoduration->jsonschema[format-nongpl]>=4.18.0->jupyter-
events>=0.11.0->jupyter-server<3,>=2.4.0->jupyterlab->jupyter) (1.3.0)
Requirement already satisfied: types-python-dateutil>=2.8.10 in
./.conda/lib/python3.11/site-packages (from
arrow>=0.15.0->isoduration->jsonschema[format-nongpl]>=4.18.0->jupyter-
events>=0.11.0->jupyter-server<3,>=2.4.0->jupyterlab->jupyter) (2.9.0.20241206)
Note: you may need to restart the kernel to use updated packages.
```

[60]:
```
import nltk

#   K                  NLTK
nltk.download("punkt")  # Ensure punkt is downloaded
nltk.download("punkt_tab")
nltk.download("averaged_perceptron_tagger")  # Required for some tokenizers
nltk.download("wordnet")  # Sometimes required for further NLP tasks
nltk.download("omw-1.4")  # For WordNet support
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/ioanniskoutsoukis/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package punkt_tab to
[nltk_data]     /Users/ioanniskoutsoukis/nltk_data…
[nltk_data]   Package punkt_tab is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /Users/ioanniskoutsoukis/nltk_data…
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/ioanniskoutsoukis/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     /Users/ioanniskoutsoukis/nltk_data…
```

```
[nltk_data]   Package omw-1.4 is already up-to-date!
```

[60]:
```
True
```

1.3 B 2: A /Φ A

[61]:
```python
# Φ
file_path = "wsj_untokenized.txt"

# Δ
with open(file_path, "r", encoding="utf-8") as f:
    text = f.read()

# Π            500
print(text[:500])

# Π              500
print(text[-500:])
```

```
 Pierre Vinken, 61 years old, will join the board as a nonexecutive director
Nov. 29. Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC,
was named a nonexecutive director of this British industrial conglomerate. A
form of asbestos once used to make Kent cigarette filters has caused a high
percentage of cancer deaths among a group of workers exposed to it more than 30
years ago, researchers reported
ivil War, and ''all have shared the view that such lawmaking power is beyond the
reach'' of the president. Sen. Kennedy said in a separate statement that he
supports legislation to give the president line-item veto power, but that it
would be a ''reckless course of action'' for President Bush to claim the
authority without congressional approval. Trinity Industries Inc. said it
reached a preliminary agreement to sell 500 railcar platforms to Trailer Train
Co. of Chicago. Terms weren't disclosed.
```

1.4 B 3: Tokenization NLTK, spaCy BERT

Σ , tokenization .

1.4.1 Π

Γ tokenizer, :

- **NLTK (`word_tokenize`)**
  - B `Punkt Sentence Tokenizer`.
  - X tokens.
  - Δ (/).
- **spaCy (`en_core_web_sm`)**
  - M **A** .
  - Π (tokens).

  - Δ tokens .
- **BERT Tokenizer (`bert-base-cased`)**
  - X **subword tokenization** (Byte-Pair Encoding).
  - E .
  - Δ .
  - Π tokens **-** (. ., `"playing"` → `["play", "##ing"]`).

[62]:
```
import nltk
import os
import spacy
from transformers import BertTokenizer
from nltk.tokenize import word_tokenize


nltk_data_path = 'os.path.expanduser("~/nltk_data")'
nltk.data.path.append(nltk_data_path)

# Re-download punkt to path
nltk.download("punkt", download_dir=nltk_data_path)

#   Tokenization   NLTK
tokens_nltk = word_tokenize(text)

#   Tokenization   spaCy
nlp = spacy.load("en_core_web_sm")
tokens_spacy = [token.text for token in nlp(text)]

#   Tokenization   BERT
bert_tokenizer = BertTokenizer.from_pretrained("bert-base-cased")
tokens_bert = bert_tokenizer.tokenize(text)

# Π           tokens
print("  NLTK Tokens:", tokens_nltk[:10])
print("  spaCy Tokens:", tokens_spacy[:10])
print("  BERT Tokens:", tokens_bert[:10])
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     os.path.expanduser("~/nltk_data")…
[nltk_data]   Package punkt is already up-to-date!
```

```
 NLTK Tokens: ['Pierre', 'Vinken', ',', '61', 'years', 'old', ',', 'will',
'join', 'the']
 spaCy Tokens: [' ', 'Pierre', 'Vinken', ',', '61', 'years', 'old', ',',
'will', 'join']
 BERT Tokens: ['Pierre', 'Vin', '##ken', ',', '61', 'years', 'old', ',',
'will', 'join']
```

1.5 B 4: Σ Tokenization (E 1)

Σ , tokenization.

1.5.1 Π Υ

- **#tokens** → T tokens .
- **#types** → T .
- **TTR (Type-Token Ratio)** → O `#types / #tokens`, .
- **Hapax Legomena** → Λ .
- **Hapax Dislegomena** → Λ .

H tokenizer N Zipf.

[63]:
```
from collections import Counter
import pandas as pd
from IPython.display import display

# Σ
def compute_stats(tokens):
    token_count = len(tokens)  # Σ     tokens
    type_count = len(set(tokens))  # M     tokens
    freq = Counter(tokens)  # Σ

    hapax_legomena = sum(1 for count in freq.values() if count == 1)  # Λ     ␣
 ↪       1
    hapax_dislegomena = sum(1 for count in freq.values() if count == 2)  # Λ  ␣
 ↪         2
    ttr = type_count / token_count  # Type-Token Ratio

    return token_count, type_count, ttr, hapax_legomena, hapax_dislegomena

# Υ            tokenizer
stats_nltk = compute_stats(tokens_nltk)
stats_spacy = compute_stats(tokens_spacy)
stats_bert = compute_stats(tokens_bert)

# Δ      DataFrame
df = pd.DataFrame(
    [stats_nltk, stats_spacy, stats_bert],
    columns=["#tokens", "#types", "TTR", "Hapax Legomena", "Hapax Dislegomena"],
    index=["NLTK", "spaCy", "BERT"]
)

df_transposed = df.T
df_transposed = df_transposed.applymap(lambda x: f"{x:.4f}".rstrip('0').
 ↪rstrip('.') if isinstance(x, float) else x)

display(df_transposed)
```

```
/var/folders/lf/_ckhk2j55jnbpz483z1xl58m0000gn/T/ipykernel_82585/2828264578.py:3
0: FutureWarning:

DataFrame.applymap has been deprecated. Use DataFrame.map instead.


                      NLTK    spaCy    BERT
#tokens              93530    95894  112325
#types               12000    11477   10266
TTR                 0.1283   0.1197  0.0914
Hapax Legomena        6254     5746    3851
Hapax Dislegomena     1830     1790    1724
```

1.6 B 5: Σ Tokenization

A tokenization :

- **Π Tokens:**
  - T **BERT** tokens, subwords.
  - T **spaCy** tokens NLTK.
- **Π Unique Types:**
  - T **NLTK** , .
  - T **BERT** subword tokenization.
- **Type-Token Ratio (TTR):**
  - **NLTK** .
  - **BERT** TTR, subwords.
- **Hapax Legomena / Dislegomena:**
  - **NLTK** hapax legomena.
  - **BERT** , subwords.
  - T **N** **Zipf**, 1-2 .

Σ : T NLTK , BERT tokens , spaCy .

1.7 B 6: Σ Tokens M T Π (E 2)

Σ , **15** tokens tokenization.

1.7.1 Σ

- N 3 .

- N   **tokenizer** “   ”   .
- N   **BERT tokenizer**   **subwords**.

[64]:
```
import random
from nltk.tokenize import word_tokenize, sent_tokenize

#  Δ
sentences = sent_tokenize(text)

#  E                              15
random_sentence = next(s for s in sentences if len(s.split()) >= 15)

#  Tokenization   NLTK
tokens_nltk = word_tokenize(random_sentence)

#  Tokenization   spaCy (                    )
tokens_spacy = [token.text for token in nlp(random_sentence)]

#  Tokenization   BERT (               tokenizer)
tokens_bert = bert_tokenizer.tokenize(random_sentence)

# Π
print("  E      Π    :\n", random_sentence)
print("\n  Tokens    NLTK:", tokens_nltk)
print("\n  Tokens    spaCy:", tokens_spacy)
print("\n  Tokens    BERT:", tokens_bert)
```

```
  E      Π    :
  Pierre Vinken, 61 years old, will join the board as a nonexecutive director
Nov. 29.

  Tokens    NLTK: ['Pierre', 'Vinken', ',', '61', 'years', 'old', ',', 'will',
'join', 'the', 'board', 'as', 'a', 'nonexecutive', 'director', 'Nov.', '29',
'.']

  Tokens    spaCy: [' ', 'Pierre', 'Vinken', ',', '61', 'years', 'old', ',',
'will', 'join', 'the', 'board', 'as', 'a', 'nonexecutive', 'director', 'Nov.',
'29', '.']

  Tokens    BERT: ['Pierre', 'Vin', '##ken', ',', '61', 'years', 'old', ',',
'will', 'join', 'the', 'board', 'as', 'a', 'none', '##xe', '##cut', '##ive',
'director', 'Nov', '.', '29', '.']
```

1.8 Σ   Σ   Tokens

A   tokens   ,   :

- **NLTK** (`word_tokenize`)

- Δ (. ,) tokens.
- **Δ** (. . "Pierre Vinken" tokens).
- T "Nov." token.

- spaCy (en_core_web_sm)
  - **Δ** **"Pierre Vinken"** .
  - E (' ') , .
  - Δ tokens.
- **BERT (bert-base-cased)**
  - **Δ** **subwords**:
    - * "Vinken" → ["Vin", "##ken"]
    - * "nonexecutive" → ["none", "##xe", "##cut", "##ive"]
  - Δ .
  - T tokens.

1.8.1 Σ

K **tokenizers** **"Pierre Vinken"** .
T BERT **subwords,** (. .).
T NLTK **spaCy** , .
A , **tokenizer** , **Named Entity Recognition (NER),**
.

1.9 A Σ Tokens M Tokenization (E 3)

Σ , tokens (**types**)
30% tokenization.

1.9.1 Σ

- T types .
- E types **30%** **tokens**.
- Σ : **Π types** **3** **tokenization;**

[65]:
```
from collections import Counter
import pandas as pd
from IPython.display import display

# Σ                         types           30%     tokens
```

```
def get_top_30_percent(tokens):
    freq = Counter(tokens)  # K                     tokens
    sorted_freq = sorted(freq.items(), key=lambda x: x[1], reverse=True)  #␣
 ↪T
    total_tokens = len(tokens)  # Σ         tokens
    threshold = 0.3 * total_tokens  # Ο       30%    tokens

    top_types = []
    cumulative_count = 0

    for token, count in sorted_freq:
        top_types.append((token, count))
        cumulative_count += count
        if cumulative_count >= threshold:
            break  # Σ                30%         tokens

    return top_types

#  B              types           tokenization
top_nltk = get_top_30_percent(tokens_nltk)
top_spacy = get_top_30_percent(tokens_spacy)
top_bert = get_top_30_percent(tokens_bert)

#  M
nltk_set = set(token for token, _ in top_nltk)
spacy_set = set(token for token, _ in top_spacy)
bert_set = set(token for token, _ in top_bert)

#  B           tokens       3
common_tokens = nltk_set & spacy_set & bert_set

#  Δ      DataFrame                           unique column names
df_nltk = pd.DataFrame(top_nltk, columns=["Token", "Σ     NLTK"]).
 ↪set_index("Token")
df_spacy = pd.DataFrame(top_spacy, columns=["Token", "Σ     spaCy"]).
 ↪set_index("Token")
df_bert = pd.DataFrame(top_bert, columns=["Token", "Σ     BERT"]).
 ↪set_index("Token")

#  E
df_final = df_nltk.join(df_spacy, how="outer").join(df_bert, how="outer")

#  E
print(" Σ     Tokens & Σ     Tokenization:")
display(df_final)

print("\n K   Types    3 M    :")
```

```
df_common = pd.DataFrame(list(common_tokens), columns=["K    Types"])
display(df_common)

#  E              tokens
print(f"\n Π        types      3      : {len(common_tokens)}")
```

```
 Σ     Tokens & Σ     Tokenization:

        Σ     NLTK  Σ      spaCy  Σ     BERT
Token
                   NaN              1.0             NaN
##ken              NaN              NaN             1.0
,                  2.0              2.0             2.0
.                  NaN              NaN             2.0
61                 1.0              1.0             NaN
Pierre             1.0              1.0             1.0
Vin                NaN              NaN             1.0
Vinken             1.0              1.0             NaN
years              1.0              NaN             NaN


 K    Types     3  M     :

  K    Types
0      Pierre
1           ,


 Π        types       3      : 2
```

1.10 Σ Π K N Zipf (E 4)

Σ , N Zipf,
:

P(r) A / r

Ό : - P(r): / token r
- r: (1 , 2 , 3 …)
- A:

E **A = 0.1** **1.0** .
Σ **Zipf** ,
.

```
[69]: import numpy as np
      import plotly.graph_objects as go
      from collections import Counter

      #  P
# Tokenization
# Tokenization   NLTK (          )
tokens_nltk_tot = word_tokenize(text)

# Tokenization   spaCy (                    )
tokens_spacy_tot = [token.text for token in nlp(text)]

# Tokenization   BERT (                tokenizer)
tokens_bert_tot = bert_tokenizer.tokenize(text)


# Π
freqs_nltk = Counter(tokens_nltk_tot)
sorted_freqs = sorted(freqs_nltk.values(), reverse=True)

#max_rank = 50
tokens = tokens_nltk_tot  # Η tokens_spacy_tot, tokens_bert_tot

#  Π
freqs = Counter(tokens)
freqs_prob = np.array(sorted_freqs) / sum(sorted_freqs)
sorted_freqs_graph = sorted(freqs_prob, reverse=True)#[:max_rank]
ranks = np.arange(1, len(sorted_freqs) + 1)

#  Δ
fig = go.Figure()

#  Π
fig.add_trace(go.Scatter(
    x=ranks,
    y=sorted_freqs_graph,
    mode='markers',
    name="Π       Σ     ",
    line=dict(color='black', width=3),
    text=[f"Rank: {r}, Freq: {f:0.2f}" for r, f in zip(ranks,␣
 ↪sorted_freqs_graph)],
    hoverinfo="text"
))

#  Τ   Α    0.1    0.1
A_values = np.arange(0.1, 1.1, 0.1)  #               1.0

#  Κ     Zipf
for A in A_values:
    zipf = A / ranks
    text = [f"A: {A:.1f}<br>Rank: {r}<br>Zipf Prob: {z:.2f}" for r, z in␣
 ↪zip(ranks, zipf)]

    fig.add_trace(go.Scatter(
        x=ranks,
        y=zipf,
        mode='lines',
        name=f"Zipf A = {A:.1f}",
        line=dict(dash='dash', width=1),
        text=text,
        hoverinfo='text'
    ))


#  P
fig.update_layout(
    title="  K     Zipf                A (            )",
    xaxis=dict(
        title="Θ   (Rank)",
        type="log",
        tickvals=[1, 10, 100, 1000, 10000, 100000],
       # ticktext=['1', '2', '5', '10', '20', '50'],
    ),
    yaxis=dict(
        title="Π    E     ",
        type="log",
        tickvals=[100, 200, 500, 1000, 2000, 5000, 10000],
        ticktext=['100', '200', '500', '1K', '2K', '5K', '10K']
    ),
    width=1080,
    height=720,
    legend_title="K     Zipf"
)

fig.show()
```

1.11 Σ Zipf

- H N Zipf, (ranks > 15).
- ϒ « » (5–10).
- H $A = 1.0$ A

1.12 X ChatGPT N Zipf (E 5)

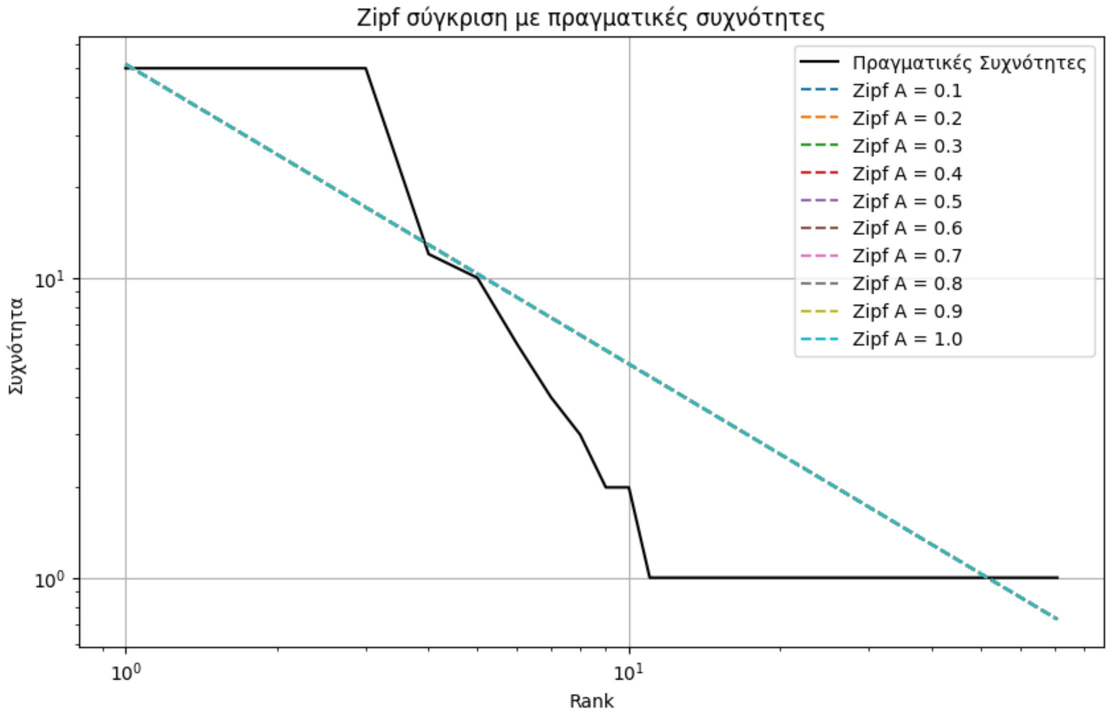
1.12.1 Prompt :

> «Δ Python N Zipf , A=0.1 1.0, log-log .»

1.12.2 K ChatGPT:

[47]:

```python
import numpy as np
import matplotlib.pyplot as plt
from collections import Counter


# Δ    :
if isinstance(text, list):
    text = " ".join(text)

words = text.lower().split()

# Υ
freqs = Counter(words)
sorted_freqs = sorted(freqs.values(), reverse=True)
ranks = np.arange(1, len(sorted_freqs) + 1)

# Σ
plt.figure(figsize=(10, 6))
plt.loglog(ranks, sorted_freqs, label="Π       Σ    ", color='black')

A_values = np.arange(0.1, 1.1, 0.1)
for A in A_values:
    zipf = A / ranks
    zipf *= sum(sorted_freqs) / sum(zipf)   #
    plt.loglog(ranks, zipf, linestyle='--', label=f"Zipf A = {A:.1f}")

plt.xlabel("Rank")
plt.ylabel("Σ    ")
plt.legend()
plt.grid(True)
plt.title("Zipf                     ")
plt.show()
```



1.12.3 Π A K ChatGPT

1. **Π Tokenization**
   - Χ `text.split()` `nltk.word_tokenize()`, `spaCy` BERT.
   - Α , , , . .
2. **Τ string**
   - Α `text` (. . tokenizer split), `.split()` (`AttributeError`).
   - Δ .
3. **Α**
   - Δ .
   - Η (. . `zipf *= sum(sorted_freqs) / sum(zipf)`) .
4. **Κ Zipf A**
   - Ο A (0.1 1.0) , .
   - Δ A.
5. **Α**
   - Τ `matplotlib`, (. . hover, tooltips, zoom).
   - Δ (rank).
6. **Ο**
   - Τ log-log Zipf .
   - Δ .

1.12.4 E A (Δ)

Σ , ,
Zipf

[68]:
```
import numpy as np
import plotly.graph_objects as go
from collections import Counter
from nltk.tokenize import word_tokenize

# Σ                         (      ranks          )
def get_zipf_data(tokens, max_rank=None):
    freqs = Counter(tokens)
    sorted_freqs = sorted(freqs.values(), reverse=True)
    if max_rank:
        sorted_freqs = sorted_freqs[:max_rank]
    ranks = np.arange(1, len(sorted_freqs) + 1)
    return ranks, np.array(sorted_freqs)

def generate_visual():

    # Tokenization
    # Tokenization   NLTK (          )
    tokens_nltk_tot = word_tokenize(text)

    # Tokenization   spaCy (                      )
    tokens_spacy_tot = [token.text for token in nlp(text)]

    # Tokenization   BERT (                tokenizer)
    tokens_bert_tot = bert_tokenizer.tokenize(text)


    # Π
    freqs_nltk = Counter(tokens_nltk_tot)

    sorted_freqs = sorted(freqs_nltk.values(), reverse=True)
    best_A = 1 # sorted_freqs[0]
    max_rank = 5000000


    # Π         tokens
    print("Π    tokens   NLTK:", len(tokens_nltk_tot))
    print("Π    tokens   spaCy:", len(tokens_spacy_tot))
    print("Π    tokens   BERT:", len(tokens_bert_tot))

    # Δ           tokenizer
    ranks_nltk, freqs_nltk = get_zipf_data(tokens_nltk_tot, max_rank)
    ranks_spacy, freqs_spacy = get_zipf_data(tokens_spacy_tot, max_rank)
    ranks_bert, freqs_bert = get_zipf_data(tokens_bert_tot, max_rank)

    total_tokens = len(tokens_nltk_tot)  # Ħ token_spacy
    zipf_pred = best_A / ranks_nltk
    zipf_pred = zipf_pred * (sum(sorted_freqs[:max_rank]) / sum(zipf_pred))


    # Δ
    fig = go.Figure()

    # NLTK
    fig.add_trace(go.Scatter(
        x=ranks_nltk, y=freqs_nltk, mode='lines+markers',
        name='NLTK',
        line=dict(shape="spline"),
        text=[f'Rank: {r}, Freq: {int(p)}' for r, p in zip(ranks_nltk,␣
↪freqs_nltk)],
        hoverinfo='text'
    ))

    # spaCy
    fig.add_trace(go.Scatter(
        x=ranks_spacy, y=freqs_spacy, mode='lines+markers',
        name='spaCy',
        line=dict(shape="spline"),
        text=[f'Rank: {r}, Freq: {int(p)}' for r, p in zip(ranks_spacy,␣
↪freqs_spacy)],
        hoverinfo='text'
    ))

    # BERT
    fig.add_trace(go.Scatter(
        x=ranks_bert, y=freqs_bert, mode='lines+markers',
        name='BERT',
        line=dict(shape="spline"),
        text=[f'Rank: {r}, Freq: {int(p)}' for r, p in zip(ranks_bert,␣
↪freqs_bert)],
        hoverinfo='text'
    ))

    # Zipf
    fig.add_trace(go.Scatter(
        x=ranks_nltk, y=zipf_pred, mode='lines',
        name=f'N   Zipf (A = {best_A})',
        line=dict(dash='dash', color='black'),
        text=[f'Zipf Predicted (A={best_A})<br>Rank: {r}, Freq: {int(p)}' for␣
↪r, p in zip(ranks_nltk, zipf_pred)],
        hoverinfo='text'
    ))

    # P
    fig.update_layout(
        title=" Σ     K          N     Zipf (  -       )",
        xaxis=dict(
            title='Θ   (Rank)',
            type='log',
            tickvals=[1, 10, 100, 1000, 10000, 100000],
            #ticktext=['1', '2', '5', '10', '20', '50'],
            showgrid=True,
            gridcolor='LightGray'
        ),
        yaxis=dict(
            title='Π    E     ',
            type='log',
            tickvals=[100, 200, 500, 1000, 2000, 5000, 10000],
            ticktext=['100', '200', '500', '1K', '2K', '5K', '10K'],
            showgrid=True,
            gridcolor='LightGray'
        ),
        hovermode='closest',
        legend=dict(
            title='M    Tokenization',
            x=1,
            xanchor='right',
            y=1
        ),
        plot_bgcolor='white',
        autosize=False,
        width=1280,
        height=720
    )



    fig.show()

generate_visual()
```

```
Π   tokens    NLTK: 93530
Π   tokens    spaCy: 95894
Π   tokens    BERT: 112325
```