Aplikacja webowa – **maks. 15 pkt**:

- Biblioteka Django / Flask / FastAPI - 15 pkt.
- Biblioteka Dash / Streamlit – 10 pkt.

Wielkość i przechowywanie analizowanego zbioru danych – **maks. 20 pkt**:[1]

- 100 – 999 rekordów/dokumentów – 3 pkt.
- 1.000 – 9.999 rekordów/dokumentów – 5 pkt.
- 10.000 – 25.000 rekordów/dokumentów – 8 pkt.
- Ponad 25.000 rekordów/dokumentów– 10 pkt.
- Przechowywanie wykorzystywanego zbioru danych w relacyjnych bazach danych (np. SQL Lite) – 10 pkt.

Dodatkowe unikatowe cechy w zbiorze danych dla rekordów – **maks. 10 pkt** (0,5 pkt za każdą od jedenastej).

Różnorodność zasobów z danymi (np. stat.gov.pl, Wikipedia, Wikidane, DBpedia, OpenAlex etc.) – **maks. 15 pkt** (5 pkt za każde unikatowe źródło).

Polepszenie jakości danych poprzez łączenie z różnych pól/źródeł (w tym różnych tabel z bazy danych) – **maks. 10 pkt** (1 pkt za każdą polepszoną unikatową cechę/pole).

Możliwość wykorzystania filtra (poprzez listę rozwijaną, przycisk opcji, przełączniki etc.) w wynikach wyszukiwania – **maks. 10 pkt** (1 pkt za każdy filtr).

Generowanie/opisywanie wyników w postaci unikatowego tekstu - **maks. 15 pkt**:

- Na podstawie reguł, analizy tekstu – maks. 10 pkt.
- Z użyciem gotowych modeli językowych (np. API do wybranego LLM) – maks. 5 pkt.
- Z użyciem własnych modeli językowych („wytrenowanych” na własnych zbiorach) – maks. 15 pkt.

Parsowanie stron HTML – **maks. 20 pkt**:

- Z użyciem wyrażeń regularnych – maks. 15 pkt (1 pkt za każdy wykorzystany wzór).
- Z użyciem specjalnych bibliotek do parsowania HTML (np. Beautiful Soup) – maks. 10 pkt. (2 pkt za każdą regułę).

Nadawanie wag termom – **maks. 25 pkt**:

- TF-IDF - 10 pkt
- Wykorzystanie osadzeń (embeddings) – maks 10 pkt.
- Zapis wektorów – 5 pkt.
- Użycie technologii FAISS – 10 pkt.

Podobieństwo termów – **maks 12 pkt**:

- Odległość Levenshteina (np. w celu poprawy błędów) – 5 pkt
- Podobieństwo k-gramów (indeks Jaccarda) – 7 pkt

---

[1] **Uwaga**: minimalna liczba unikatowych cech dla rekordów w zbiorze danych – co najmniej 10. Każda cecha musi być wykorzystana przez system.

Relewancja i miary efektywności (wyszukiwarki) – **maks. 10 pkt**:

- Użycie miar precyzji, pełności etc. – maks. 10 pkt.

Miary podobieństwa (dokumentów/zapytań) – **maks. 20 pkt**:

- Miara iloczynu, Dice'a, Jaccarda, cosinusa, MMR – maks. 10 pkt. (5 pkt za każdy).
- Użycie LSI, ColBERT – 10 pkt.

Obliczanie ważności dokumentów - **maks. 10 pkt**:

- PageRank – maks. 10 pkt

Klasyfikacja dokumentów (w oparciu o teksty/termy/metadane) – **maks. 30 pkt**:

- Naiwny klasyfikator Bayesa (Multinomial Naive Bayes) – maks.10 pkt
- Model Bernoulliego (Bernoulli Naive Bayes) – maks.10 pkt
- Klasyfikacja Rocchio – maks. 10 pkt.
- k-najbliższych sąsiadów – maks. 10 pkt
- SVM – maks. 10 pkt
- Las Losowy, Gradient Boosting – maks. 10 pkt
- Wytrenowanie własnego modelu językowego do klasyfikacji (fine-tuning encoderów typu BERT, SBERT) – 20 pkt.
    - Wykorzystanie ponad 5 klas/kategorii/etykiet – maks. 10 pkt (0,5 pkt za każdą dodatkową kategorię).
    - Należy przedstawić informacje na temat ewaluacji modelu.

Analiza tekstu – **maks. 40 pkt**:

- Podstawowa morfologiczna - maks. 10 pkt.
- Identyfikacja nazw własnych – maks. 10 pkt.
- Analiza sentymentu - maks. 10 pkt
- Wielojęzyczność analizowanych tekstów – maks. 20 pkt (5 pkt. za każdy dodatkowy (nie podstawowy) język).
- Ekstrakcja informacji z użyciem API GenAI – 5 pkt.
- Klasyfikacja z użyciem API GenAI (LangChain lub LlamaIndex) – 5 pkt.

Zamieszczenie wykresów (z wynikami) – **maks. 20 pkt**:

- Za każdy unikatowy typ wykresu (np. liniowy, słupkowy, radarowy, chmura słów, mapa cieplna etc.) – 4 pkt za każdy.
- Interaktywny wykres – 2 pkt za każdy

Zapisywanie wyników bieżących aplikacji do pamięci podręcznej (cache) – **5 pkt**.

Uwaga! Końcowa liczba punktów zależy od liczby osób w zespole na podstawie wzoru:

$$\frac{\text{otrzymanePunkty}}{\text{liczbaOsób}}$$