**Cel projektu**

Glowne cele projektu to:

1. Zapoznanie sie z dobrymi praktykami przy pracy nad jednym projektem w zespole
   1. korzystanie z systemu kontroli wersji (git)
   2. prowadzenie dokumentacji (mediawiki)
   3. oprogramowanie do zarzadzania projektem (zadania, bledy; Redmine)
2. Zapoznanie sie z podstawowymi pojeciami i algorytmami zwiazanymi z literature-mining.
   1. Jako zrodlo danych posluza pelne teksty artykulow pobrane z Pubmed Central w formacie xml
3. Stworzenie prostego algorytmu do wyszukiwania korelacji miedzy dowolnym slowem(query) a slowami w publikacjach bazy Pubmed Central

**Etapy projektu**

1. Przygotowanie srodowiska (git, mediawiki, Redmine)
2. Stworzenie parser'a dla publikacji w formacie xml
   1. kazdy paragraf jako pojedynczy rekord
   2. usuniecie [stopwords](http://www.netautopsy.org/umlsstop.htm)
   3. usuniecie innych niepotrzebnych string'ow (np. znacznikow )
   4. [stemming](http://tartarus.org/%7Emartin/PorterStemmer/index.html)
3. Stworzenie trzech baz danych
   1. unikalny rekord paragrafu vs slowa w nim wystepujace
   2. dowolne slowo vs lista identyfikatorow paragrafow w ktorych wystepuje
   3. dowolne slowo vs liczba jego wystapien we wszystkich dostepnych w bazie Pubmed Central publikacjach
4. Implementacja prostego scenariusza wyszukiwania korelacji dla danego slowa/slow(query) stanowiacego zapytanie
   1. wykonanie zapytania w bazie danych (zapytanie sql;jeden rekord dla kazdego slowa)
   2. znalezienie paragrafow z danym slowem/slowami (zapytanie sql; jeden rekord dla kazdego paragrafu)
   3. stworzenie statystyki slow dla wszystkich paragrafow zawierajacych dane slowo/slowa
   4. wyszukanie globalnej czestosci wystepowania slow znalezionych w poprzednim punkcie (zapytania sql; jeden rekord dla kazdego slowa)
   5. analiza korelacji pomiedzy slowem-query a znalezionymi slowami w bazie (oparta na z-score); na bazie czestosci wystepowania znalezionych slow w calej bazie Pubmed Central (background probability) i w zestawie paragrafow zawierajacych slowo-query
5. Opracowanie przykladowego zapytania