EasyTranslate

LDA topic model

task1: design a prototype hierarchical LDA topic model using English Wikipedia as corpus based on Genesis's Python framework https://radimrehurek.com/gensim/,
task2: retrieve distributions over topics to infer top level document categories for the source texts
task3: sort source documents according to high level categories
task4: implement a Python interface to compare topic distributions in the source document space
task5: retrieve similar source documents based on a distance measure
task6: model linguistic domain skills based on topic distributions of translated source documents

IBM Watson
compare the above LDA topic modeling against source document descriptions retrieved using state of the art natural language processing functionalities based on the IBM Watson services https://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/ :

alchemy
task7: named entity extraction (english german french 100+ types, disambiguation),
task8:keyword extraction (english german french ranked relevance, sentiment), hierarchical taxonomy (1000 english topic task8: categories, confidence scores),
task9:concepts (implicit extraction of english high-level abstractions, ranked relevance, linked data),
task10: document sentiment (english german french entity keyword level analysis using negation and modifiers)
task11: targeted sentiment (specified target e.g. brand),
task 12: relations (english sentence parsing)
task 13: language (detection)

natural language classifier
task14: take prepared training data (match class labels to representative example texts) to train a classifier based on wikipedia using deep learning to provide an API interface that returns classes (inferred categories) for texts it has not seen in training

concept insights
task15: go beyond text matching to provide links between words in input documents and relevant content e.g. articles topics people or related concepts based on english wikipedia

personality insights
task 16: personality analysis based on an input text e.g. retrieved from linkedin CVs or twitter messages (english, spanish, min 3500 words) to output attributes (trained on 66 LIWC psycholinguistic dictionary categories based on social media) that correlate with big5 psychological personality models (average or raw scores; 5 major traits, 30 facets) needs (10 marketing related aspects) and values (10 motivational aspects), that may be used to market brands towards individual consumers, anticipate customer needs and preferences or match personalities to corporate values