11.04 erster Kontakt zu Herrn Mendel von Lukas Baur per Email über Themen Vorschläge ihrer - und unsererseits. Dort kam von Herrn Mendel schon der Vorschlag für das OSM WIKI.

12.04 Email zur ersten Terminvereinbarung

17.04 erstes Treffen mit Herrn Mendel und besprechen mit verschiedenen Themen darunter auch das OSM Thema.

18.04 Email von Herrn Mendel mit verschiedenen Textanalyse Programmen, unteranderem das Phyton- Tool „gensim“

22.04 Email an Herrn Mendel mit weiteren Recherchen zum Thema Textanalyse

23.04 wird sind Idioten……..

24.04 Neuer Termin mit Herrn Mendel am 26.04 um 13.30 ausgemacht um weitere Schritte zu besprechen, gefundenes und mögliche Aufteilung besprechen.

26.04 Treffen mit Herrn Mendel bei dem weitere Details überlegt wurden außerdem Aushändigung des schriftlichen Vertrages für das Projekt und die Details dazu. (siehe Email von Mendel)

03.05 weitere Recherchen und Analyse des Wikis von 2013 dabei wurden schon json Dateien gefunden sowie eine Tag-Liste. Allerdings würde festgestellt das die Daten so nicht lokal zu bekommen sind.

15.05 das Team schlägt einen neuen Termin vor um weiteres Vorgehen zu besprechen und bisherige Ergebnisse zu vorzustellen.

16.05 Terminvorschlag von Herrn Mendel auf den 17.05 um 13.30

17.05 Treffen mit Herrn Mendel bei dem nun erste konkrete Aufgaben verteilt wurden.

1. Web-Crawler bauen um die bekannten Wiki-Seiten herunterzuladen hierbei hat Herrn Mendel gesagt wir sollen darauf achten dies in einem angenehmen Tempo zu machen um den Betreiber nicht zu überlasten und zu nerven.
2. Des Weiteren mussten die einzelnen Strukturen noch genauer untersucht werden um die verschiedenen Seiten zu analysieren damit einen besseren Überblick über die Seiten zubekommen in Bezug auf Aufbau und Art.
3. Falls die Daten einen heterogenen Aufbau hätten, würde hier ein Konzept zu Verbindung zwischen Tags und Sprache angebracht. Sinnvoll wäre hierbei sich von Überschrift zu Überschrift zu bewegen.
4. Test-Framework für das Programm umsetzen mit integrierter Architektur für Anfrage-Funktion.

27.05 – 10. 06 Die einzelnen Aufgaben wurden in den folgenden Wochen abgearbeitet.

Zu 1. Wir haben zum weiteren Vorgehen direkt 2 Crawler gebaut um auch sicher an die Daten zu gelangen ….. So wurde ein Cralwer für die „KEYS“ gebaut und einen für die „TAGS“ was dazu führt das wir leicht die kompletten Wiki-Seiten crawlen konnten die wir wollten. Schlussendlich wurde entschieden das ganze über die Tag-Liste zu machen.

Zu 2. Nach gründlichen Untersuchungen wurde festgestellt, dass die Seiten einen sehr heterogenen Aufbau haben und es somit nicht so einfach war die Seiten einfach alle gleich zu behandeln und zu verwerten. Allerdings gab es ein paar Sachen die doch öfters vorkamen wie unteranderem eine Sidebar, diverse Tabellen und einige Überschriften die sich auf unterschiedlichen Seiten wiederholten wie z.B. „See also“ oder „How to use“.

3. Da die Seiten nicht sehr homogen waren hat sich dieser Punkt leider erledigt und man musste eine andere Lösung finden. Wir entschieden uns dies über … ( hier fehlt text)

4. Nach unseren ersten Vorstellungen ergab sich als Beispiel eine Black Box welche auf eine Anfrage eine definierte Tag-Übersicht zurückgibt, was auf einem REST-Server laufen sollte. Allerdings wurde zum dem Zeitpunkt nicht weitere daran gearbeitet.

14.06 erneutes Treffen mit Herrn Mendel da von den oberen Aufgaben einiges erledigt wurde und man die neuen Fortschritte präsentieren wollte um weiteres Vorgehen zu besprechen. In den nächsten Schritten sollte die Wiki-HTML Seiten weiter aufbereitet werden so auch Tabellen erkannt werden und auch Tags in der Umgebung zu finden sind. Außerdem sollten gewisse Textteile vor und nach dem Tag mit dem Tag möglichst in Verbindung gebracht werden. Dadurch sollte ein Mapping von Wörtern auf relevante Tags erreicht werden.

18.06 es wurde beschlossen ein json-Format zu erstellen, dass dann nur über die einzelnen Seiten als Maske gelegt wird um die nötigen Daten zu erhalten. Allerdings stellte auch hier sich anfangs das Problem das die sehr unterschiedlichen Seiten alles Mögliche enthielten und HTML Seiten ein sehr bescheidene Aufbau hatten, da keine strukturierten Container verwendet wurden sondern nur Überschrift dann Text, Überschrift dann Text, …. . Aufgrund dieser Tatsache wurde dann entschieden nur die am öftesten verwendeten Überschriften samt ihres Inhaltes zu benutzen. Dafür wurde ein Programm geschrieben das alle Überschriften sammelt und absteigend nach Anzahl sortiert. Wir sind dann zu dem Schluss gekommen, dass die ersten 30 davon ausreichend sein sollten, da alle danach nur noch insgesamt 20-mal vorkamen auf allen Seiten was sehr spezifisch war. Außerdem wurde festgestellt das auch bei den Überschriften keine Homogenität besteht so wurden verschiedene Überschriften gefunden mit dem gleichwertigen Inhalt so z.B. bei „See also“ und „see also“.

21.06 Nach weiteren Diskussionen wurde dann doch festgestellt das es eigentlich nur sinnvoll ist ausgewählte Überschriften selber rauszusuchen.

28.06 Es wurde das erste Mal ein konkretes json-Format erstellt. (Beispiel einfügen)

30.06 Es wurden weitere Fortschritte im Bereich des json-Format gemacht und die dabei ausgelesen Daten waren auch einigen Maßen schön anzuschauen um damit weiter zu arbeiten. Es wurde ein weiteres Treffen mit Herrn Mendel vereinbart.

02.05 Diskussion über Gewichtung der einzelnen Überschriften wie „See also“ -> sehr wichtig oder „Problems“ -> sehr unwichtig. So wurde überlegt ein dreistufen Verfahren einzuführen welches mit wichtig, normal und ignorieren gestaltet wird. Außerdem wurde inzwischen versucht mit dem TF-idf-Maß die Relevanz der einzelnen Terme zu bewerten. Des Weiteren stellte sich die Überlegung ob wir eine Tag zu Tag Beziehung brauchen und wie diese am besten eingesetzt werden würde.

-> Klären beim nächsten TREFFEN..