Meeting agenda

• Good software **engineering practices** for evaluation of the whole project.

• How to better implement extreme programming.

• Discussion about the upcoming report.

Summary of Final Product Design

• First and foremost, we would like to discuss about our final product. We believe we have a more palpable understanding of it as a product. First of all, it will be composed of two different packages, one is designed with the user in mind, the other with the developer.

• User-centered design:

Seek will run in the cloud, allow the user to upload data in multiple formats (see list of features) in order to have Seek search and index it, to query Seek, and to train Seek. Both come with a very friendly, futuristic user interface (see interface). This is designed for a speech-based interaction system between the user and Seek, but also for the use of the keyboard and mouse to actually collect, visualise and query the data should you be in a more professional environment.

• Developer-oriented design:

Seek will be available as a python package to download and install to Linux or Mac OS (sorry, Windows). Developers are encouraged to build modules that would allow Seek to spread its skills,

We are envisioning an open-source knowledge platform for Seek’s developers and teachers, but this is not covered in our current project.

*To discuss: Is the double-package model viable? Should we mention the far-fetched design ideas in the final project (i.e. speech-based interaction, knowledge platform)?*

Summary of past milestone - MILESTONE 1 (week 2 - week 7)

• PHASE 1: Gatherer

• crawl and collect files containing textual data from websites (i.e. Spiral)

• extract text from multiple file types

• (IN PROGRESS) collecting and processing textual information from XML (i.e. Wikipedia)

• PHASE 2: Executor

• script to setup everything

• set up server

• run batch commands for many files

• test and time various large tasks

• (IN PROGRESS) process data in the cloud with the cloud installation

• (IN PROGRESS) allow the user to choose between using Seek locally, in the cloud, or both

• (IN PROGRESS) connect to cloud from local

• PHASE 3: Topic Modelling

• PHASE 4: Seek: mind and body

• PHASE 5: Training Seek

To discuss: Organising remaining phases and tasks. Do we narrow the user base and make it slightly specialist (i.e. only finance, only history, only geography? maybe use exactly the recipe Herox want, and only if there is time do all the poetry)

Summary of our work

• As we have just hit mid-November, we believe we are nearly finishing the first time-boxed iteration of our timeline. During this first milestone, we have achieved the following tasks:

• All PHASE 1, everything urgent and which can’t be worked on in parallel from PHASE 2

• 2 weeks initial research and accommodation period: learned to work in pairs, switch pairs, communicate efficiently: frequent and useful commits; we have finished in advance doing various ML courses and tutorials; distributed conceptual tasks/chose gurus; we have researched means of performing further layers of learning after the Topic Modelling

• We have collected all the necessary dependencies for the initial data-collection and data conversion phase, and created a first working draft of the setup script for the Developer package of which we make use of now

• We are successfully collecting text from multiple sources and formats, and converting it to plain text, both with concurrent methods allowing to batch data - more than enough for both functionality and large enough knowledge resources for now.

• We are already planning and preparing for PHASE 3: Statistician: parsing has been attempted and looks promising, thanks to Marc Deisenroth’s advice to try some libraries for topic modelling

• We have fully designed the user interface, and created all the necessary graphics (PHASE 4)

• We have set up a cloud server, we will integrate it in the current work during the Second Milestone

• and added the following features

• Gatherer 0.1.0 : converts pdf to plain text

• Gatherer 0.1.1 : scrapes Spiral repo for PDF

• Gatherer 0.1.2: converts text from .DOC, .ODT, .XLS etc

• Gatherer 0.1.3: converts text from simple images

• Gatherer 0.1.4:

• Executor 0.1.0 : basic interface for handing files over to the Gatherer

• Executor 0.1.1: extended for triggering batch conversions from any format available so far to plain text

• Executor 0.2.0: Concurrent Executor 0.1

• Executor 0.2.1 (IN PROGRESS): more interaction and control (flags for commands etc)

• Executor 0.2.2 (IN PROGRESS): client-server communication

• Statistician 0.1.0 (IN PROGRESS): performs parsing for topic modelling and basic analysis