*Trước khi làm thì phải update trên server về, vào phần work này để biết đang làm gì !*

*Có 2 file : Tiến, Cường mỗi người làm trên file này sau đó tổng hợp lại sau*

--------------------------------------------------------------------------------------------------------------------------------------------------------12/8----------------------------------------------------------------------------------------------------------------------------------------------------------------------

**+ Công Việc đặt ra:**

- Hoàn thành Crawl.

- Tìm hiểu google scholar API.

- Thu thập tài liệu để làm tài liệu training của ngành CS <trước hết là theo cái khung >

- Sử dụng SVN lib.

- Đọc các tài liệu có trong Document, tài liệu đọc rồi thì cho ra phần đã đọc

+ Tóm tắt lại kiến thức có trong tài liệu.

**+ Một số thư viện số có trên mạng cần tìm hiểu**

ACM <http://portal.acm.org/dl.cfm>

IEEE Computer Society <http://www.ieeexplore.ieee.org/Xplore/guesthome.jsp>

Association for Computational Linguistics <http://www.aclweb.org/>

Springer-Verlag <http://www.springer.com/>

Citeseer : <http://citeseerx.ist.psu.edu/>

The DBLP Computer Science Bibliography : <http://www.informatik.uni-trier.de/~ley/db/>

**+ Ý kiến và câu hỏi của Thầy đóng góp cho nhóm :**

- Xác định chính xác khung phân loại (Tiến có đề xuất theo cái wiki đó đọc file khung phân loại cs ý ), tức là mình phải tìm hiểu một số thư viện số như IEE,ACM rằng những hệ thống đó có cho ta tìm kiếm theo chủ đề, topic không ? Nếu có thì những chủ đề đó nó phân chia như thế nào ,có nhiều cấp không ---> từ đó hoàn thiện cái khung của mình.

- Thu thập các bài báo (một số lượng đủ lớn ) để ta làm training và test thử sau đó .

- Tìm hiểu rõ ràng và cài đặt thử hoặc có thể là dùng lại socuercode của 1 phương pháp phân loại (SVM) phải chi tiết từng bước như là đánh chủ mục làm sao, rút đặc trưng, đánh trọng số làm sao.

- Xây dựng kiến trúc của chương trình ? Trên desktop or web ?

- Các thư viện số có phân loại không ? Nếu có thì phân loại như thế nào ? Lấy được thông tin gì trong web đó về tài liệu (abtract , tile , metadata ….? ) .

- Xác định lại mục tiêu của đề tài ? phân loại để làm gì ? tại sao phải crawl trong khi đó các thư viện có thể nó không cho download tài liệu về và các thư viện số đó đã phân loại rồi có cần thiết phải phân loại không ??????? <gợi ý là crawl các thông tin về bài báo như abtract để phân loại làm dữ liệu làm giàu ontology dùng để lưu vết các bài báo cung cấp các thông tin …… , >.

**Công việc làm :**

* Xây dựng kiến trúc của chương trình.
* Xem các thư viện điện tử được liệt kê ở trên và viết tóm tắt về các thông tin gồm :

+ Có cho tìm kiếm theo chủ đề,tác giả …

+ Khung phân loại của nó nếu có.

+ Thông tin của 1 bài báo lấy được gì ? metadata,abtract, …

Tiến : ACM,Citier,DBLP.

==🡺 Chỉnh sửa lại đề cương !

--------------------------------------------------------------------------------------------------------------------------------------------------------16/8----------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Việc đang làm :**

* Khảo sát các thư viện số .
* Tìm một số bài báo để làm key trong đề tài.
* Chỉnh sửa đề cương.
* Xem crawl có thu được file HTML để rút ra abtract của một bài báo.
* Suy nghĩ kiến trúc của chương trình (các modul, Package, class ….).

**Góp ý của Thầy :**

* Phân lớp tài liệu trong lĩnh vực CS dựa trên phần Abtract của mỗi bài báo .

**=🡺 tìm hiểu các paser để mình lấy các phần abtract của bài báo trong file PDF,HTML,DOC,…**

* Suy nghĩ về việc dính cái tìm kiếm bài báo về (Lấy phần abtract) .
* Dùng Crawl

Tool trong đó cho phép lấy và search thông tin về bài báo từ các thư viện điện tử : <http://jabref.sourceforge.net/index.php>

Phân loại download và phần lấy abtract dựa trên cái gói này chương trình này làm .