Problem - Classify web pages using the SMCFL algorithm.
https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14582

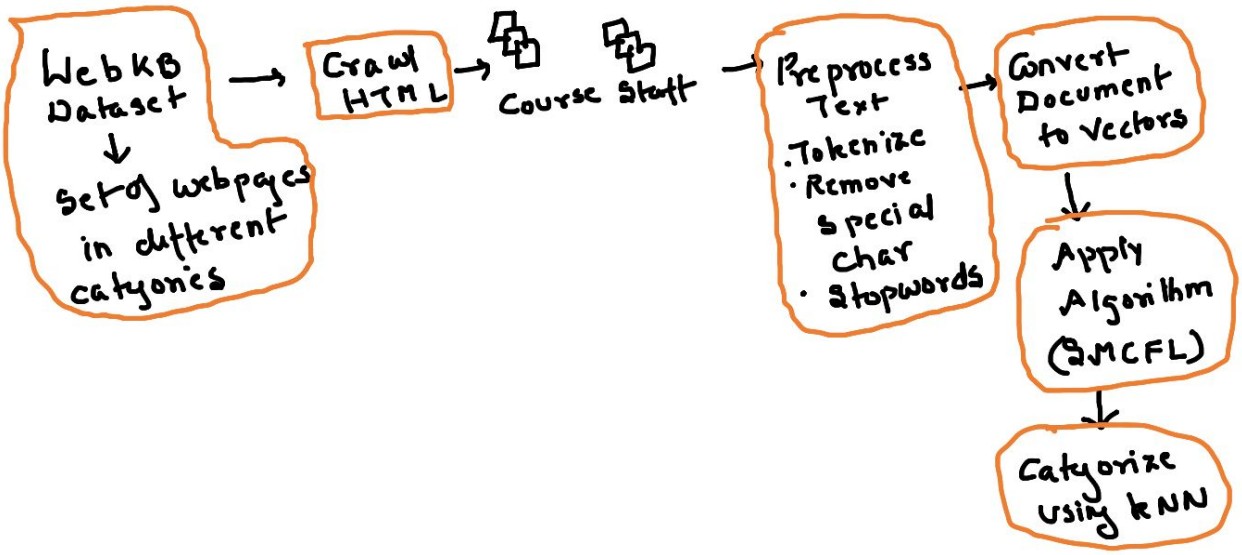
Dataset - WebKB Dataset.
http://www.cs.cmu.edu/~webkb/

Background -

Webpage data is often multi-view and high-dimensional, and the webpage classification application is usually semi-supervised.

Due to these characteristics, using semi-supervised multi-view feature learning (SMFL) technique to deal with the webpage classification problem has recently received much attention.

Multi-view learning is concerned with the problem of machine learning from data represented by multiple distinct feature sets. Like in web-page classification, a web page can be described by the document text itself and at the same time by the anchor text attached to hyperlinks pointing to this page.

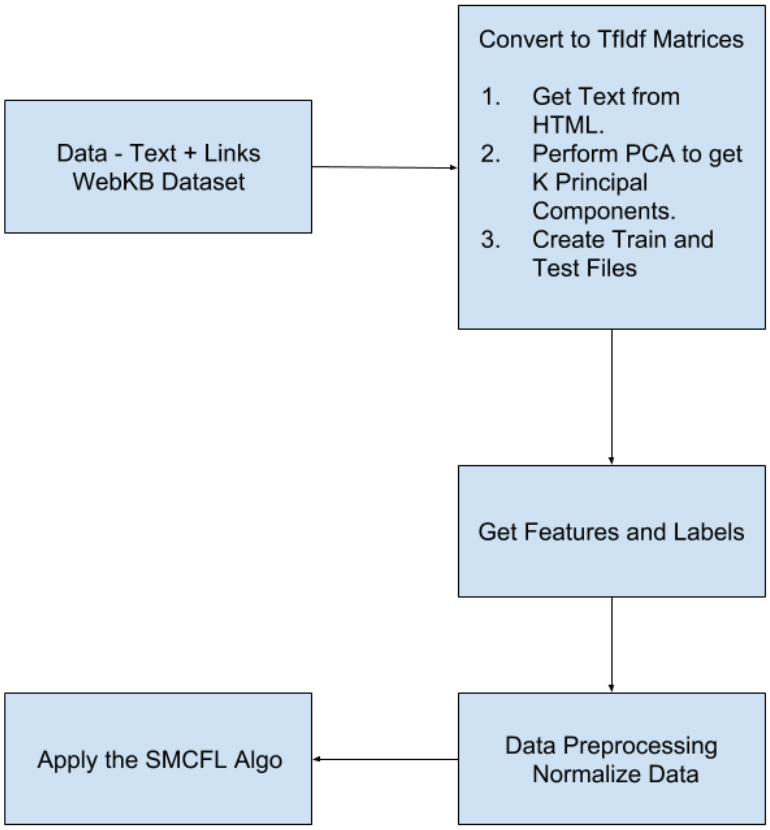


Algo -

Input: Training sample sets X l and X u , test sample y .
Output: Class label of y.

Step 1. Calculate 'a' (alpha) according to (20).
Step 2. Calculate H according to (17).
Step 3. Calculate W according to (22) or (23).
Step 4. Obtain the projected test sample Z^y and the projected

labeled training sample set Z ^X .
Step 5. Use the nearest neighbor classifier with the cosine distance to classify Z^y according to Z^X .



How to effectively utilize the correlation information among multi-view of webpage data is an important research area.
Correlation analysis on multi-view data can facilitate extraction of the complementary information.

Webpage classification has three characteristics:

(1) Webpage is a kind of multi-view data since it usually contains two or more types of data, e.g.,text, hyperlinks and images, where each type of data can be regarded as a view. These multiple views describe the same webpage.
(2) Webpage classification is a semi-supervised application, since labeled pages are harder to collect compared to unlabeled pages in practice.
(3) Webpage data is high-dimensional, since webpages usually contain much information.

Considering these three characteristics, it is crucial to design effective semi-supervised multi-view feature learning (SMFL) methods.

To our knowledge, two webpage classification methods taking these three characteristics into account have been developed, namely semi-paired and semi-supervised generalized correlation analysis (SSGCA) and uncorrelated semi-supervised intra-view and interview manifold discriminant (USI2MD).

Motivation and Contribution
The correlation information from inter-view and intra-view depicts association relation among multiple views which has close connection with classification.
Intra View means relationship b/w samples within a certain view. Inter View means samples across different views.

Observation : Take 10 webpage samples -> perform PCA to get 2 major components. Upon plotting it is observed

Samples of different views hold small corelation.

Between class samples within each view hold larger corelation.

We should maximize the corl of samples b/w diff views, while in same class while minimize cor of samples b/w same views, while from diff classes. To make distribution more favorable to classification.

It is necessary to perform cor analysis on webpage data which can help us learn features with favorable separability.

Currently most of SMFL methods do not consider cor info. For few there exists room to improve their discriminat capabilities.

Most of current webpage classification methods do not consider all three characteristics of webpage classification.

Summary of the new approach-
1. SMFCL - The objective function is designed to maximize cor b/w intra class samples and minimize inter class cor.
2. We transform matrix variable based nonconvex objective function into a convex quadratic programming problem with 1 real var. The soln is global optimal and can be derived analytically without iterative calculation.
3. It's verified on two widely used webpage datasets and can outperform state of art webplage classification methods.