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

Dataset
We evaluate our approach on WebKB Dataset. It's a 4 Universities Data Set. The WebKB dataset contains 1051 webpages from two classes (230 pages in the course class and 821 pages in the non-course class). Each webpage is characterized by the page view and the link view. 8,282 pages manually classified into student, faculty, staff, courses etc.

Background
Webpage data is often multi-view and high-dimensional, and the webpage classification application is usually semisupervised.
Due to these characteristics, using semisupervised multi-view feature learning (SMFL) technique to deal with the webpage classification problem has recently received much attention.

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. 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.
2. Webpage classification is a semisupervised 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.

Correlation Analysis
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.

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 correlation.
- Between class samples within each view hold larger correlation.
- In same class we should maximize the corl of samples b/w diff views and in different class minimize cor of samples b/w same views. To make distribution more favorable to classification.

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

Most of current webpage classification methods do not consider all three characteristics of webpage classification.
Currently most of SMFL methods do not consider cor info. For few there exists room to improve their discriminant capabilities.

Summary of the new approach
1. SMFCL - The objective function is designed to maximize correlation b/w intra class samples and minimize inter class correlation.
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.

Semi-supervised Multi-view Correlation Feature Learning (SMCFL)

Objective Function
Suppose that X ^ l = {X1 X2 Xc} is the labeled training webpage sample set from C classes.
Let X^u be the the unlabeled training sample set.
Let N denote the total sample number in X
We define the within-class correlation S-w , between-class correlation S-b and total correlation S-t as follows: See eqns 1,2,3 in paper.

max (w) f(w) = S-w - r1 * S-b - r2 * S-t where r1 and r2 >0 and known as weight coefficients.

min(H) 1/2 * | H | ^ 2 + n * e where n is regularization and e is slack variable to relax constraint.

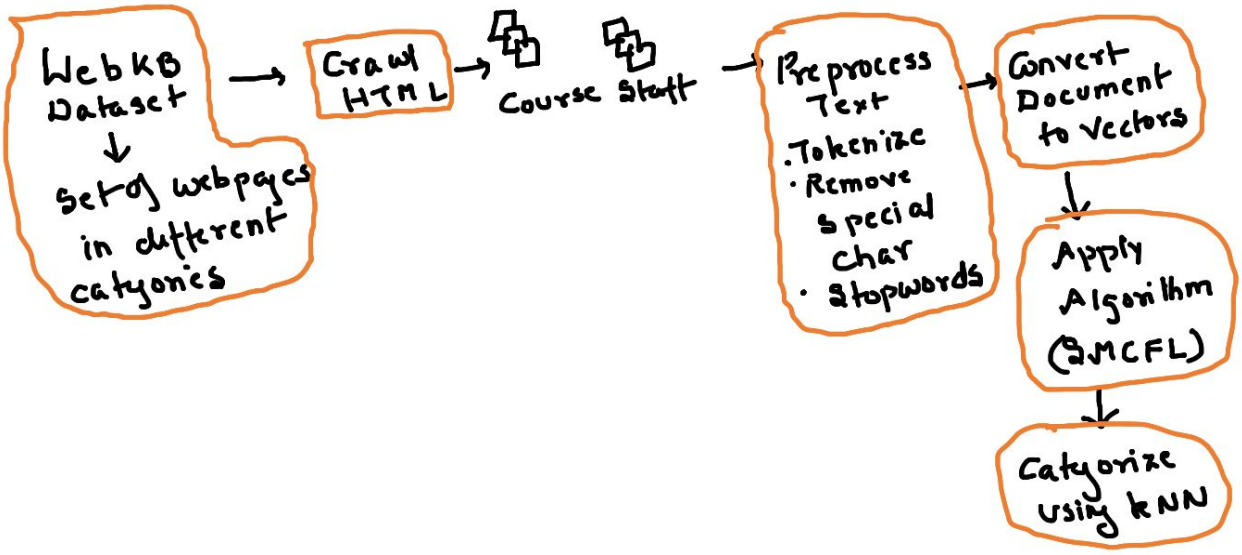
By applying the Lagrangian technique to constrained optimization problem, we define the Lagrange function as eqn 13

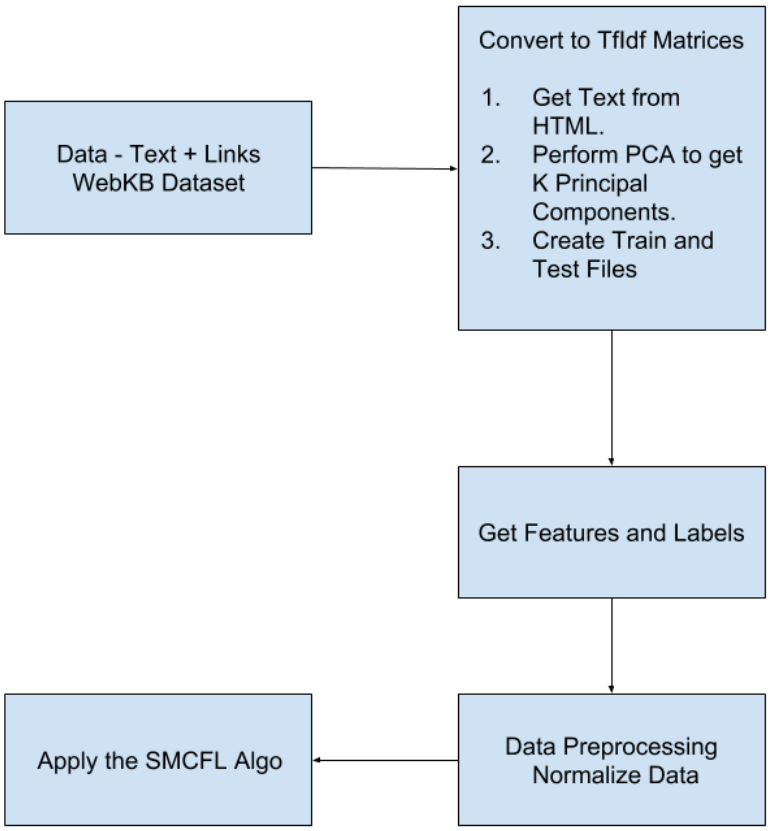
Algorithm

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 .

$Z(s) \hat{} X = W^T * X(s)$
$Z(s) \hat{} y = W^T * y(s)$

Design





References

https://github.com/rnjtsh/Semi-Supervised-MultiView-Feature-Learning : Project Impl
http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/webkb-data.gtar.gz : Dataset
http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/ : Dataset Home Page
https://www.researchgate.net/profile/Shiliang_Sun2/publication/257436121_A_survey_of_multi-view_machine_learning/links/5a66b7600f7e9b6b8fde5659/A-survey-of-multi-view-machine-learning.pdf
https://github.com/niyasc/Webpage-classification-Minor-Project
https://india.endurance.com/machine-learning-website-categorization/ : A wonderful article that explains the steps to classify websites.