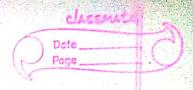


Scanned with CamScanner

6	Date Page
	Processed. while stop words typically rely to the most common wand, there is no universal list of stop words.
	OSTEMMING > for grammitical reasons, don are going to use diff. forms of a work (e.g. crosnized, organizing, etc). Additionally
	there are families of denivitavelly relative words with similar meaning like democratization.  The goal of sterming and lemmatization
	is to neduce inflectional forms & sometime derivationally linked forms of a world to its common base form
	con, one, is -> be  con, cons, cons, con's -> con  -> when applied to a doc, result will be
	Somewhat like this  the boy's cons one diff. colon  the by con be differ colon
	a heuristic that chops off the end of a word in the hope of achieving the
	goal connectly more of time, and after includes hemoval of derivational afficient
	readure selection is a Process of Selecting a subset of the towns occurring



in the memaining set and using only this Subset as fecture in text classification. 7 St somes 2 loopuses. final, it makes training & applying a classifier more efficient by decreasing the size of vicab. this is of Penticular importance to classifica that are expensive to triain. - Second, feature Selection often increases clossification accuracy by elementing raise teatre these features when added to does representati increases the dossification emon on new date. -> Vectorization is a Brocess of converting text into machine gredable from weards are repriesented as (Vectors (nurronically) -> (ount vectorizen (one-not entading) involves counting the number of occurrence for each world occasing in the doc. -> Simple, vector is seated having as many dimentions as there are distinct words in text/doc/collection of dues being used. each unique word has a unique alimention & is represented by a 1 in the dimention with Gs everywhere ese. This results in huge & sparse vectors that capture no melational data.

classmate -> If IDF rectore one related to one m encoded vectors, but instead of just feating a count, they feature runerical repres -ation, where trong words aren't just bress on not Present instead, they are suppress by their lean foreg. multiplied by invoy Loc. Heg. - In simpler terms, wands that accom everywhere should be given very lit brounde a longe and - of value, Heur if a word opposes very little on leg frequently but only in Specific Place, then they are Rossible of higher Significance -> the down side is there to no capture of Senantic relatedness, this is Solved with the 100-convence mate neural Probablitistic Condusion: × Text does were tokenized fillered, vectorized, Dothus Successfully Bracessed basics of MLP were understood