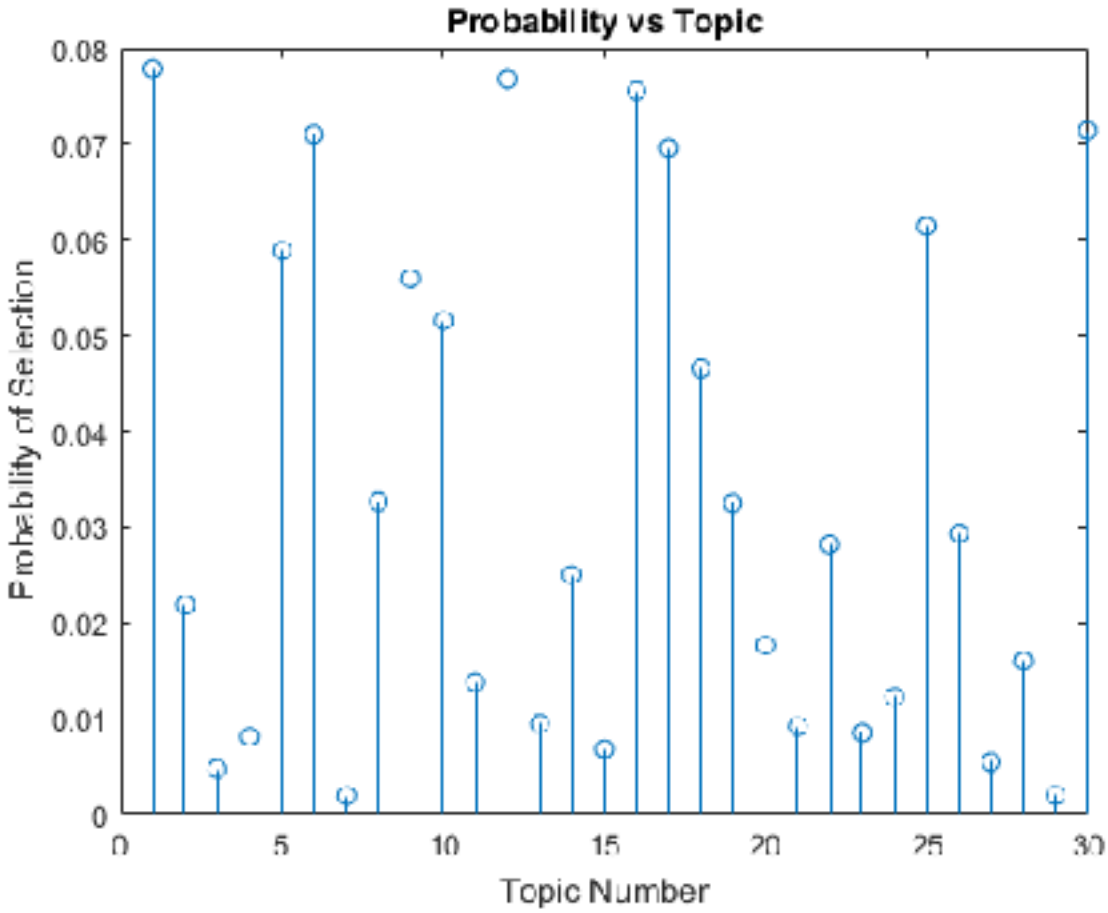
CS 498 AML: Homework 6

Rauhul Varma (rvarma2), Tadas Aleksonis (alekson2)

EM Topic models

The UCI Machine Learning dataset repository hosts several datasets recording word counts for documents here. You will use the NIPS dataset. You will find (a) a table of word counts per document and (b) a vocabulary list for this dataset at the link. You must implement the multinomial mixture of topics model, lectured in class. Cluster this to 30 topics, using a simple mixture of multinomial topic model, as lectured in class.

a. Produce a graph showing, for each topic, the probability with which the topic is selected.



b. Produce a table showing, for each topic, the 10 words with the highest probability for that topic.

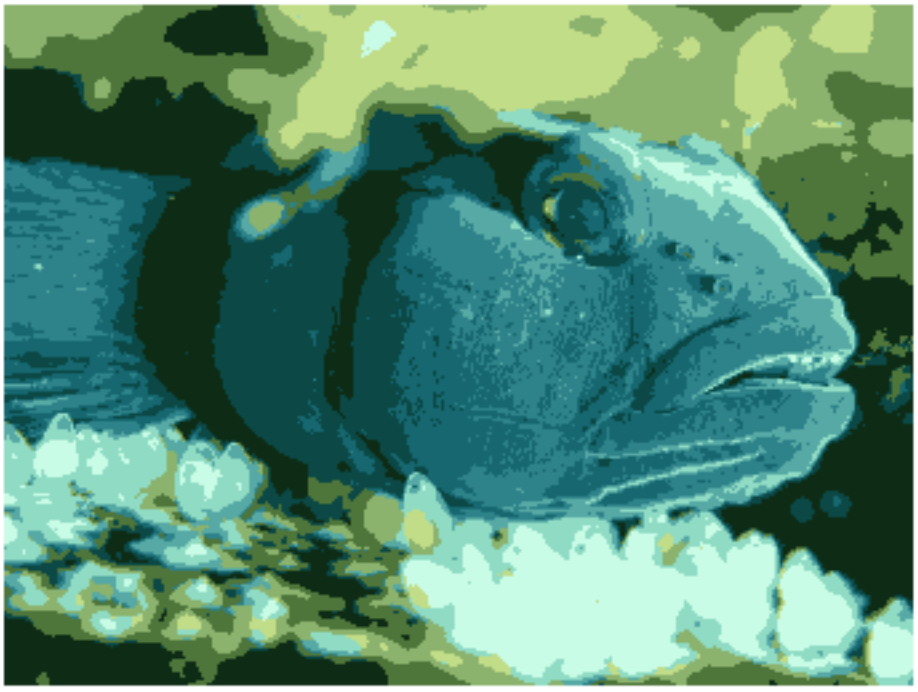
model	cell	input	neuron	visual	unit	network	direction	motion	field
neuron	network	model	neural	input	system	cell	pattern	spike	function
model	control	learning	network	system	neural	controller	movement	dynamic	forward
function	network	neural	threshold	input	weight	bound	set	number	result
target	speaker	system	location	display	cue	classifier	learning	cues	data
network	unit	input	learning	training	string	set	neural	weight	grammar
network	chip	weight	neuron	neural	learning	input	analog	system	output
model	data	network	algorithm	parameter	set	distributi	learning	problem	tree
learning	function	action	algorithm	model	policy	problem	system	control	optimal
network	set	algorithm	training	model	predisposit	data	method	exemplar	component
component	algorithm	function	learning	data	network	informatio	signal	matrix	input
network	model	speech	system	training	recognition	neural	set	data	word
model	data	gaussian	algorithm	network	parameter	function	distribut	set	input
function	algorithm	set	learning	model	data	problem	vector	error	training
network	error	learning	weight	training	function	set	input	generalizat	neural
network	model	disparity	digit	rotation	output	option	activatio	unit	input
network	unit	input	neural	learning	output	system	training	weight	hidden
model	learning	movement	network	neuron	apg	part	neural	motor	system
network	model	unit	variables	response	gaussian	conditiona	density	structure	markov
input	voltage	chip	wta	signal	analog	offset	output	circuit	neural
unit	input	vector	space	network	hidden	data	algorithm	weight	learning
learning	routing	network	reinforceme	path	load	control	node	packet	traffic
image	network	model	object	images	set	learning	system	point	neural
input	parameter	neuron	potassium	output	model	conductanc	calcium	dendrite	rate
word	network	input	model	pattern	recognition	set	character	vector	training
cell	orientati	input	model	cortical	neuron	visual	response	contrast	cortex
tuning	populatio	width	encoding	neuron	information	fisher	neural	field	stimulus
circuit	output	system	input	current	analog	model	chip	signal	network
network	training	set	input	error	learning	data	neural	classifier	pattern
input	parameter	neuron	potassium	output	model	conductanc	calcium	dendrite	rate

Image segmentation using EM

You can segment an image using a clustering method - each segment is the cluster center to which a pixel belongs. In this exercise, you will represent an image pixel by its r, g, and b values (so use color images!). Use the EM algorithm applied to the mixture of normal distribution model lectured in class to cluster image pixels, then segment the image by mapping each pixel to the cluster center with the highest value of the posterior probability for that pixel.

a. Segment each of the test images to 10, 20, and 50 segments. You should display these segmented images as images, where each pixel's color is replaced with the mean color of the closest segment

Fish 10 segments



Fish 20 segments

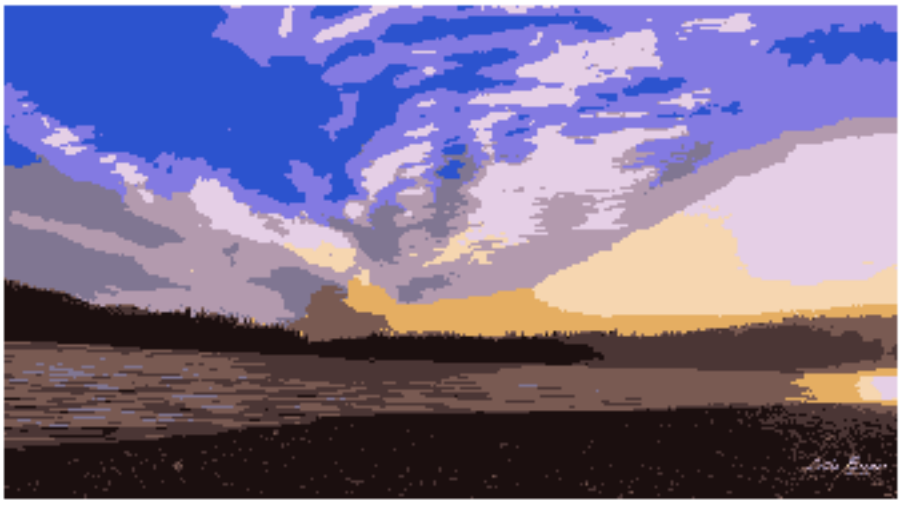
Fish 50 segments

Flower 10 segments

Flower 20 segments

Flower 50 segments

Sky 10 segment



Sky 20 segments
> See part b for all 5 versions of this segmentation

Sky 50 segments

b. We will identify one special test image. You should segment this to 20 segments using five different start points, and display the result for each case. Is there much variation in the result?

   There is virtually no variation in the 5 segmented images. The images are below: