Hierarchical Text Categorization for MALACH

J. Scott Olsson
olsson@math.umd.edu

w/ Doug Oard of UMIACS, CLIS
Multilingual Access to Large spoken ArCHives

- UMD, JHU, IBM, Charles University, the University of West Bohemia, the Visual History Foundation
- Many areas of study (automatic speech transcription, machine translation, cross-language information retrieval, text categorization, etc.)

Goal: “to improve access to large multilingual collections of recorded speech in oral history archives”
The Shoah VHF Dataset

Audiovideo testimonies of Holocaust survivors

Unique features: size, coherency

- 51,649 testimonies, roughly 2.5 hours each, in 32 languages.
- The largest, most complex, coherent digital video library on Earth (180+ Tb).

This differs significantly from other available datasets.
Automatic Text Categorization
A Flat Model: $k$ Nearest Neighbors

A Hierarchical Model: HPLC

Two Models

Auschwitz II

Concentration Camps

Auschwitz

Auschwitz II
Measuring Success

Binary measures of success:
• accuracy
• precision, recall
• F-measure

We can do better: hierarchical measures

How close is our guess to the correct answer?
Questions?

olsson@math.umd.edu

http://www.math.umd.edu/~olsson/hiercat/