ML @ Snackable

Alexander Tkachenko
Outline

• What we do and why
• Spoken language
• ML project lifecycle
Spoken Word Content

- Interviews
- Social video
- Radio
- YouTube
- Podcasts
- Audiobooks
- Speeches
- TV
Snackable AI Engine

Spoken Word Content

Speech recognition
Speaker Detection

Transcript
Speakers
Chapters
Chapter titles
Paragraphs
Summaries
Snacks
Metadata
Ads
Music
Search
Building the habits that you need to get results

James Clear

And so throughout the rest of this presentation, I want to talk about how we can do that. Today, I’m going to teach you how to build the habits that you need to get the results that you want. And in order to do this, I’m going to take you through a framework for building better habits. And then I’m also going to share a personal example of how I use this. So my writing habit. If you don’t know, I write a James Clear dot com. I write about how to build better habits, improve performance, and generally live better. Over a million people visit the site each month. There’s over 400,000 subscribers on the weekly email newsletter, and it all came out of the simple writing habit.

So for the rest of this talk, there are four stages. Have information. I’m going to take you through each of those four. All right. So the four stages are noticing, wanting, doing and liking, noticing, wanting, doing and liking. You cannot perform a habit or take an action if you do not notice something. I need to see a coffee cup sitting on the side in order to pick it up.
Snacks

And one way to think about it is just kind of basic math. Like if you just look at the numbers, if you were able to improve by 1% each day for an entire year and those gains compound, you would end up 37 times better at the end of the year. And if you were to get 1% worse, you would whittle yourself almost all the way down to zero. And what we what’s interesting here is that everybody wants a transformation, right? Everybody wants a radical improvement on rapid success. But we fail to realize that small habits and little choices are transforming us every day already, that these times when you make a choice is slightly better, slightly worse, a little mistake or a small error 1% better or 1% worse that these things compound over time. Good habits are the compound interest of self-improvement. And so if you can learn to master those, then you can make time work for you rather than against you. Right? Good habits make time your ally. Bad habits make time your enemy.
Use Cases

• Promotion: sharing snacks / audiograms in social media
• Rich transcript pages
• Improved navigation: chapters on YouTube
• Search
Spoken Language

• Less structured, less grammatical
• Disfluencies
  • Uh, um, word repetitions, etc.
• Multiple speakers
• Low information density
• Variety of topics
What it means

• ASR is often not perfect:
  • word transcription (especially entities)
  • sentence boundaries and punctuation.

• Models trained on written text can have poor results.
Solution

• Fix ASR using custom vocabulary
  • Analyze metadata (titles, rss, descriptions, etc)
  • Video - OCR
• Fine-tune text-based model
• Use additional signals from audio/video
  • Scene detection
  • Acoustics features
ML Project Lifecycle

Planning ➔ Data ➔ Modeling ➔ Deployment
ML Project Lifecycle

Planning
Data
Modeling
Deployment

• Define a Problem
• Data
• Resources
• KPI / Metrics
ML Course/Competition/Research

• Problem
  • is well-defined in terms of X and Y.
  • the goal is to beat a baseline.

• Data:
  • is provided by organizer

• KPIs/Metrics
  • defined by organizer

• Resources
  • personal time, computational power
In Industry

- Let’s try something cool
- High-level goals:
  - Improve a product
  - Gain competitive advantage
  - Optimize a process
- Estimate cost/risk/reward/resources
- KPIs/Metrics
  - ML metrics vs. business metrics
- ML vs. business side.
ML Project Lifecycle

Do we have data?

**No:**
- Where to get data?
  - use public data
  - buy
  - annotate in-house/outsource
- How much data is needed?

**Yes:**
- Is our data usable?
  - exploratory analysis
  - what’s missing?
- Build required sample:
  - export, clean, normalize, transform.
ML Project Lifecycle

- Planning
- Data
- Modeling
- Deployment

- Writing code
- Training model
- Error analysis
Iterative Model Development

Model, Data

Error Analysis

Training
Data-centric Approach

Research
- Model
- Error Analysis
- Training

Industry
- Data
- Error Analysis
- Training
Reproducibility

- Documentation is crucial
- Version control
  - Data
  - Code
  - Model weights
  - Performance metrics
- Tools
  - Code: Github
  - Data/metrics: dvc, github lfs, etc...
- It’s easy!
Example: Chapter Segmentation

Phase 1 – get started quickly
  • Simple unsupervised model
  • “Simple” data: news, podcasts
Model Development

• Tuning parameters:
  • Type of embeddings?
  • Text preprocessing (stopwords, lemmatization, weighting)
  • How many chapters?
  • …
Error Analysis

• Good:
  • detecting topic drift in information-dense content (e.g. news)

• Bad:
  • Fails with less informative content (e.g. conversations)
  • Can’t learn useful signals:
    • “Let’s now move to the next question…”
    • “Second, I want to talk about ….”
  • Identified topic boundaries are inexact.
Chapter Segmentation: Phase 2

- Data: podcasts, talk shows, interviews, webinars.
- More powerful supervised model.
Annotation project

- Data collection
  - Decide which types of content are important
  - Create a representative and balanced sample
- Define what is “chapter”
  - and chapter categories (ad, music, intro, outro, presentation, sub-topic, etc, …)
- Implement a simple annotation tool.
- Hire annotators.
Model development

- **Challenges**
  - Inconsistent annotations
  - Huge variation in accuracy on different data samples

- **Solution**
  - “debug” annotations using model assistance

- **Result**
  - Quickly got excellent results without tuning model.
Example 2: snack detection

Iteration 1
- no data
- unsupervised algorithm
- good enough!

Iteration 2
- some data
- simple supervised model
- much better results!

Iteration 3
- even more data
- DNN model
- great results!
Take aways

• Get started with something simple, setup a baseline.
• Unsupervised approach – easy to start, but has limitations.
• Supervised approach is worth the effort.
• Data is more important, than the model.
• Model-assisted approach helps to quickly identify bugs in data.
ML Project Lifecycle

**Planning**
- Monitoring
- Maintaining

**Data**

**Modeling**
- ML issues
  - Data drift
  - Unexpected inputs
- Software engineering issues

**Deployment**
Data Drift

• Training set:
  • podcasts, news
• Real inputs:
  • zoom meetings, webinars
Monitoring

• Model inputs and outputs
  • Trends
  • Outliers
• Client feedback
• System health
Monitor Model Inputs / Outputs

• Manual analysis
  • e.g. examine 10 random files a day

• Track metrics
  • Inputs
    • File type
    • File duration
    • Number of speakers
    • Word rate
  • Model Outputs
    • Number of chapters predicted
    • Smallest/largest chapter duration
Customer Feedback

Implicit feedback

- Editing: snack boundaries, transcript.
- User exports a snack on YouTube.
Customer Feedback

Implicit feedback
• Editing: snack boundaries, transcript.
• User exports a snack on YouTube.

Explicit feedback
• Thumbs up/down
• Surveys/interviews
Software/hardware metrics

• Server load
• Resource utilization (cpu, memory, disk, etc.)
• Processing time
  • How long it takes to process one hour of content
• Number of failed files per day.
Dashboards
Iterative Deployment

- Manual
- Automatic

- Update Model, Data
- Training
- Error Analysis

- Deploy
- Monitor
Roles in ML

• Researcher
• Data Scientist
• ML Engineer
• Data Engineer
• MLOps Engineer
Thank you!

Questions?