Categorization of Musicology Questions from Community-based Q&A Site using Latent Dirichlet Allocation

Hengyi Fu
School of Information, Florida State University
Music Stack Exchange

- An online **social Q&A site** which allows users to post questions and provide answers about **music**.
- The site was launched in **April 2011**, and until now it has **13,367** users, **27,967** questions, and **23,359** answers.

**Answer Examples**

- **Question**: How to improvise over a C#, B, A chord progression?
  - **Answer**: I'm looking to improvise a piano solo in over a C#B, A chord. I'm unsure of what scale to use for this, or unsure on how to determine why musically some notes work better than others for this progression.

- **Question**: How to model complex systems like networks in PostgreSQL?
  - **Answer**: Is there a name for a minor scale with a raised 3rd or a major scale with a lowered 6th and 7th?

**Other Features**

- Answers
- Badges
- Comments
- Events
- Privileges
- Questions
- Revisions
- Suggested Edits
- Tags
- Users
Data Collection

- Data dump published by Music StackExchange which spans **62 months**, from January 2011 to March 2016

- The final dataset used for this study contains **8,284 questions** and **18,821 answers**, after removing other types of posts (e.g. comments to questions or answers) based on post type id
Data Collection

- Data dump published by Music StackExchange which spans 62 months, from January 2011 to March 2016

- The final dataset used for this study contains 8,284 questions and 18,821 answers, after removing other types of posts (e.g. comments to questions or answers) based on post type id
Data Preprocessing

- **Preprocessing**: R *Text Mining* package and Apache OpenNLP
  - used for stop words removal, text tokenization, part-of-speech chunking

---

I have been learning how to play Spanish guitar for more than two years and half. Every time my friends ask me to play something, I get scared and my hands start to shake. How to break this fear? I know this is a subjective question as everyone has his own ways, but please share what works for you.

(a) Before pre-processing

learn play spanish guitar two year half time friend ask play scare hand start share break fear know subjective question everyone way share work

(b) After pre-processing
Latent Dirichlet allocation (LDA)

- A statistical topic modeling technique used to automatically recover topics from a corpus of text documents
- Deal with large dataset without the need for tags, training data, or predefined taxonomies

Example:

In traditional harmony, I have heard the word "sonority" used sometimes freely in a way to describe harmonic structures or harmonic systems. However, in my experience the word was used somewhat flexibly to describe any harmonic structure, express or implied, including a single stacked chord, an agglomeration of notes, and also perhaps even including other combinations of notes which, though not fitting the definition of a chord recognized in traditional harmony, would nonetheless form a system of harmony. What is the usage of the word "sonority" in the history of music, and does it have a proper definition?

Topics: Music Theory/Education: 0.92
Instrument Learning : 0.08
LDA Implementation

**Question Topic Model**
- Poster stemming algorithm
- MALLET for 500 Gibbs sampling iterations with 50 topics

**Question Type Model**
- Remove chunks that represent noun phrases
- MALLET for 500 Gibbs sampling iterations with 30 topics
Metrics

• The output of LDA is a list of *Topics, Topic Proportion* of each document/post, and *Topic Share* of each topic in the corpus

• **Topic Proportion:** what percentage of each post belongs to each topic
  – Usually a document have between one and five dominant topics
  – A threshold of 0.10 is set as the topic proportion cutoff

• **Topic Share:** A measure of how much a topic has been discussed as compared to other topics in the corpus
  – Measured by the proportion of posts that contain a certain topic
  – If a topic has a share metric of 10%, then 10% of all posts contain this topic

\[
\text{share}(z_k) = \frac{1}{|D|} \sum_{d_i \in D} \sum_{\theta(d_i, z_k) \geq \delta} \theta(d_i, z_k)
\]

where \( D \) is the set of all documents in the dataset. The *share* metric measures the proportion of documents that contain the topic \( z_k \).
## Discussion Topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>Top LDA Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>score, practice, try, problem, way, change, correct, wrong, work</td>
</tr>
<tr>
<td>T2</td>
<td>http, music, online, link, example, tutorial, read, check, pc</td>
</tr>
<tr>
<td>T3</td>
<td>piano, mode, interval, frequency, experience, scale, harmony, rhythm</td>
</tr>
<tr>
<td>T4</td>
<td>technique, keyboard, chord, exercise, tone, notation</td>
</tr>
<tr>
<td>T5</td>
<td>guitar, string, amplifier, digital-piano, keyboard, midi</td>
</tr>
<tr>
<td>T6</td>
<td>voice, perform, effects, band, improvisation</td>
</tr>
<tr>
<td>T7</td>
<td>key, train, chord, violin, inversion, finger</td>
</tr>
<tr>
<td>T8</td>
<td>sheet-music, write, element, make, example, understand, compose</td>
</tr>
<tr>
<td>T9</td>
<td>history, theory, chord-progression, classical-music, counterpoint, analysis, teach</td>
</tr>
<tr>
<td>T10</td>
<td>setup, clean, maintenance, solo, dissonance, travel, equipment</td>
</tr>
<tr>
<td>T11</td>
<td>transposition, part, set, arrange, perform, notation, dynamic</td>
</tr>
<tr>
<td>T12</td>
<td>tempo, effect, electronic, percussion, style, beat, synthesizer</td>
</tr>
</tbody>
</table>

Top twelve topics (share >=0.1) discovered by LDA
# Discussion Topics (Stack Overflow)

<table>
<thead>
<tr>
<th>Topic name</th>
<th>Top LDA words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding Style/Practice</td>
<td>code case design differ implement approach write depend example</td>
</tr>
<tr>
<td>Problem/Solution</td>
<td>work problem code solution fine issue solve work_fine wrong idea</td>
</tr>
<tr>
<td>QA and Links</td>
<td>question answer post comment article edit link example find read</td>
</tr>
<tr>
<td>Function</td>
<td>function type variable value parameter call pass return argument</td>
</tr>
<tr>
<td>Job/Experience</td>
<td>develop work time people software year dai companies product job</td>
</tr>
<tr>
<td>OO Programming</td>
<td>class method object call implement instance interface create static</td>
</tr>
<tr>
<td>UI Development</td>
<td>control view event button click set user tab create add</td>
</tr>
<tr>
<td>File Operation</td>
<td>file directories path folder create read upload open write copies</td>
</tr>
<tr>
<td>Website Design/CSS</td>
<td>css browser element div style html firefox flash work javascript</td>
</tr>
<tr>
<td>String Manipulation</td>
<td>string character encoding format space convert word quote input text</td>
</tr>
</tbody>
</table>

*Topic names are manually created to briefly describe each topic*
Major Topic Category

• **Performance-related discussion** (T1, T4, T6)
  – asking for suggestions to improve performance abilities or solve performance related issues

• **Instrument-Specific Discussion** (T3, T7, T10)
  – asking for guidance or solutions to issues for instrument learning
  – asking for suggestions tips for buying an instrument/equipment, or for instrument maintenance

• **Composition** (T8, T10)
  – sharing of knowledge and experience in composition processes (e.g. music writing, music arrangement)

• **Recording** (T5, T12)
  – discussing technical or equipment issues that occurred in the sound recording process

• **Music Theory/Education** (T9)
  – regarding a specific aspect or aspects of music theory

• **Reference Materials** (T2)
  – asking for reference sources
## Question Type/Intent

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Top LDA Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>write, to write, create, to create, compile, complete, play, to play</td>
</tr>
<tr>
<td>T2</td>
<td>doesn’t work, work, try, didn’t, won’t, isn’t, confuse, can be, would be</td>
</tr>
<tr>
<td>T3</td>
<td>learn, to learn, read, look, find, recommend, start, take, try to learn, refer, help</td>
</tr>
<tr>
<td>T4</td>
<td>wrong, lose, master, make, to make, tell, suggest, explain, understand, analyze, point</td>
</tr>
<tr>
<td>T5</td>
<td>proceed, set, count, compare, reach, make, to make, use, to use</td>
</tr>
<tr>
<td>T6</td>
<td>cause, assume, fail, work, check, to see, fine, due, notice</td>
</tr>
<tr>
<td>T7</td>
<td>to use, can do, instead, buy, configure, run, merge, order, update, create, base</td>
</tr>
</tbody>
</table>

### Seven Question Types (share >=0.1) discovered by LDA

- **Seeking Solutions**: Something has been implemented but does not work (T2, T6)
- **Seeking Instructions**: Not knowing how to implement something (T1, T5) or Not understanding how/why something works (T4)
- **Seeking Suggestions**: Suggestions for learning something (T3)
I have recently begun producing and recording some music - electronic MIDI tracks as well as analog recordings through a microphone. The DAW I am using is Samplitude Silver. I have no experience in mixing and mastering the recordings, and my attempts so far have failed to produce really listenable results. What is the best way to learn about how to mix and master the recordings to produce something that sounds well?

Seeking solutions (T6: 0.91 T5: 0.09)

I am learning piano and music theory for the first time. I got a Casio privia px 150 - it was recommended everywhere. I especially liked the weighted keys. But then today I noticed <a link> and recorded it. The clicking noise is only heard when piano is on low to medium volume and does not really affect the playing experience. If I decide to keep the piano instead of returning it, would it affect my learning experience in the long run?

Seeking suggestions (T3: 0.81 T2: 0.19)
Conclusion and Future Work

• Users’ questioning behaviors were motivated by the recognition of knowledge gaps, lack of resources, need for others’ suggestion, spanning different topics

• Users were particularly interested in finding instructions or solutions to issues that have occurred in a specific learning task

• Future research includes
  – Comparing these topic/types identified by machine-learning with expert cataloging and user-generated tags, involving tags in analysis to see the topic evolution
  – Applying LDA other textual, unstructured, music-related data (e.g. RILM datasets)
Thank you!